

The Theory of Probability

by B. V. GNEDENKO

Translated from the Russian
by GEORGE YANKOVSKY

MIR PUBLISHERS • MOSCOW

First published 1969
Second printing 1973
Third printing 1976
Fourth printing 1978

На английском языке

© English translation, Mir Publishers, 1978

Contents

Introduction	7
Chapter 1. THE CONCEPT OF PROBABILITY	13
Sec. 1. Certain, Impossible, and Random Events	13
Sec. 2. Different Approaches to the Definition of Probability	16
Sec. 3. The Sample Space	19
Sec. 4. The Classical Definition of Probability	23
Sec. 5. The Classical Definition of Probability. Examples	26
Sec. 6. Geometrical Probability	33
Sec. 7. Frequency and Probability	39
Sec. 8. An Axiomatic Construction of the Theory of Probability	45
Sec. 9. Conditional Probability and the Most Elementary Basic Formulas	51
Sec. 10. Examples	59
Exercises	67
Chapter 2. SEQUENCES OF INDEPENDENT TRIALS	70
Sec. 11. Independent Trials. Bernoulli's Formulas	70
Sec. 12. The Local Limit Theorem	76
Sec. 13. The Integral Limit Theorem	85
Sec. 14. Applications of the Integral Theorem of DeMoivre-Laplace	92
Sec. 15. Poisson's Theorem	97
Sec. 16. An Illustration of the Scheme of Independent Trials	102
Exercises	104
Chapter 3. MARKOV CHAINS	107
Sec. 17. Markov Chains Defined. Transition Matrix	107
Sec. 18. Classification of Possible States	111
Sec. 19. Theorem on Limiting Probabilities	113
Sec. 20. Generalizing the DeMoivre-Laplace Theorem to a Sequence of Chain-Dependent Trials	116
Exercises	123
Chapter 4. RANDOM VARIABLES AND DISTRIBUTION FUNCTIONS	124
Sec. 21. Basic Properties of Distribution Functions	124
Sec. 22. Continuous and Discrete Distributions	130
Sec. 23. Multidimensional Distribution Functions	134
Sec. 24. Functions of Random Variables	142
Sec. 25. The Stieltjes Integral	155
Exercises	160
Chapter 5. NUMERICAL CHARACTERISTICS OF RANDOM VARIABLES	164
Sec. 26. Mathematical Expectation	164
Sec. 27. Variance	169
Sec. 28. Theorems on Expectation and Variance	176
Sec. 29. Mathematical Expectation Defined in the Axiomatics of Kolmogorov	182
Sec. 30. Moments	185
Exercises	191
Chapter 6. THE LAW OF LARGE NUMBERS	195
Sec. 31. Mass-Scale Phenomena and the Law of Large Numbers	195

Sec. 32. Chebyshev's Form of the Law of Large Numbers	198
Sec. 33. A Necessary and Sufficient Condition for the Law of Large Numbers	206
Sec. 34. The Strong Law of Large Numbers	209
<i>Exercises</i>	218
Chapter 7. CHARACTERISTIC FUNCTIONS	219
Sec. 35. Definition and Elementary Properties of Characteristic Functions	219
Sec. 36. The Inversion Formula and the Uniqueness Theorem	224
Sec. 37. Helly's Theorems	230
Sec. 38. Limit Theorems for Characteristic Functions	235
Sec. 39. Positive Definite Functions	239
Sec. 40. Characteristic Functions of Multidimensional Random Variables	243
<i>Exercises</i>	248
Chapter 8. THE CLASSICAL LIMIT THEOREM	251
Sec. 41. Statement of the Problem	251
Sec. 42. Lyapunov's Theorem	254
Sec. 43. The Local Limit Theorem	259
<i>Exercises</i>	266
Chapter 9. THE THEORY OF INFINITELY DIVISIBLE DISTRIBUTION LAWS	267
Sec. 44. Infinitely Divisible Laws and Their Basic Properties	268
Sec. 45. The Canonical Representation of Infinitely Divisible Laws	270
Sec. 46. A Limit Theorem for Infinitely Divisible Laws	275
Sec. 47. Statement of the Problem of Limit Theorems for Sums	278
Sec. 48. Limit Theorems for Sums	279
Sec. 49. Conditions for Convergence to the Normal and Poisson Laws	282
<i>Exercises</i>	285
Chapter 10. THE THEORY OF STOCHASTIC PROCESSES	287
Sec. 50. Introductory Remarks	287
Sec. 51. The Poisson Process	291
Sec. 52. Conditional Distribution Functions and Bayes' Formula	298
Sec. 53. Generalized Markov Equation	302
Sec. 54. Continuous Stochastic Processes. Kolmogorov's Equations	303
Sec. 55. Purely Discontinuous Stochastic Processes. The Kolmogorov- Feller Equations	311
Sec. 56. Homogeneous Stochastic Processes with Independent Increments	318
Sec. 57. The Concept of a Stationary Stochastic Process. Khinchin's Theorem on the Correlation Coefficient	323
Sec. 58. The Concept of a Stochastic Integral. The Spectral Decompo- sition of Stationary Processes	331
Sec. 59. The Birkhoff-Khinchin Ergodic Theorem	334
Chapter 11. ELEMENTS OF QUEUEING THEORY	339
Sec. 60. A General Description of the Problems of the Theory	339
Sec. 61. Birth and Death Processes	346
Sec. 62. Single-Server Queueing System	355
Sec. 63. A Limit Theorem for Flows	361
Sec. 64. Elements of the Theory of Stand-by Systems	367
APPENDIX	376
BIBLIOGRAPHY	382
SUBJECT INDEX	388

Introduction

This book aims to give an exposition of the fundamentals of the *theory of probability*, a mathematical science that treats of the regularities of random phenomena.

The theory of probability originated in the middle of the seventeenth century and is associated with the names of Huygens, Pascal, Fermat, and James Bernoulli. The correspondence between Pascal and Fermat dealing with problems in games of chance that did not fit into the framework of the mathematics of those days laid the foundations for such important concepts as probability and mathematical expectation. We must be clear on one point: the famous scientists that dipped into these gambling problems also foresaw the fundamental role of the science that studies random events. They were convinced that clear-cut regularities could arise on the basis of large numbers of random events. Because of the low level of development of natural science in that period, however, games of chance and also problems of insurance and demography were for a long time the sole concrete material used in building up concepts and methods of the theory of probability. This circumstance placed its imprint on the formal mathematical apparatus applied in the solution of probability problems: it consisted exclusively in the use of elementary arithmetic and combinatorial methods. The subsequent development of probability theory and also the broad application of its results and methods of investigation to natural science, in particular to physics, demonstrate that the classical concepts and methods still hold today.

The rigorous demands of the natural sciences (the theory of errors of observation, problems in the theory of gunfire and of statistics, primarily population statistics) called for a further development of probability theory and the use of a more sophisticated analytical apparatus. A particularly important role in the development of analy-

tical methods of probability theory was played by DeMoivre, Laplace, Gauss, and Poisson. Associated with this trend, in a formal analytical manner, is the work of the founder of non-Euclidean geometry, N. I. Lobachevsky, devoted to the theory of errors in measurements performed on a sphere and carried out for the purpose of establishing the dominant geometrical system of the universe.

From the mid-19th century to the twenties of the present century, the development of probability theory was largely connected with Russian scientists: P. L. Chebyshev, A. A. Markov, and A. M. Lyapunov. This success of Russian science was prepared by the work of V. Ya. Bunyakovsky who cultivated applications of probability theory to statistics, in particular in the field of insurance and demography. He wrote the first Russian course of probability theory which exerted a profound effect in the field and excited much interest. The abiding principal value of the work of Chebyshev, Markov and Lyapunov in probability theory consists in the fact that they introduced and widely applied the concept of a random variable. In this book we shall take up Chebyshev's studies in the law of large numbers, Markov chains and Lyapunov's limit theorem.

Today, probability theory is extending its influence and practical application to many spheres, and researches throughout the world have enriched the theory with important results. In this great upsurge, the Soviet school of probability theory continues to occupy a prominent position. Foremost among the Soviet workers are S. N. Bernstein, A. N. Kolmogorov, and A. Ya. Khinchin. The ideas of and results obtained by present-day scientists that have revolutionized the theory of probability will be introduced as the subject matter demands. In the very first chapter we will speak of the fundamental studies of Bernstein and Kolmogorov dealing with the foundations of probability theory. In the first decade of this century, E. Borel pointed to ideas connecting the theory of probability with the metric theory of the functions of a real variable. Somewhat later—in the 1920s—A. Ya. Khinchin, A. N. Kolmogorov, E. E. Slutsky, P. Lévy, A. Lomnitsky and others considerably elaborated these ideas, which proved extremely fruitful for the development of science. It will be noted, for one thing, that it was precisely in this direction that a definitive solution of the classical problems posed by Chebyshev was found. The principal advances of this trend are due to J. W. Lindeberg, S. N. Bernstein, A. N. Kolmogorov, A. Ya. Khinchin, William Feller, Paul Lévy, and others. The ideas of the metric theory of functions and, subsequently, of functional analysis made it possible to extend substantially the content of probability theory. The 1930s saw the foundations laid for the theory of stochastic (probabilistic, random) processes, which today has become the principal trend of research in probability theory. This theory is a fine example of the organic synthesis of mathematical and natural-science thinking,

when the mathematician has grasped the physical essence of a crucial scientific problem and finds adequate mathematical language to fit it.

The idea of such a theory was apparently expressed by J. H. Poincaré, and the first outlines of it may be found in the work of L. Bachelier, Fokker, and Planck. However, construction of the mathematically complete foundations of the theory of stochastic processes is associated with the names of Kolmogorov and Khinchin. We must note that solutions to classical problems of probability theory were found to be closely tied up with the theory of stochastic processes. In Chapter 10 we shall give elements of this new section of probability theory. Finally, we may mention such fresh fields of application as reliability theory and queueing theory. Sections 60 to 64 of this book deal briefly with the content of this new division of science.

During recent decades the role of probability theory in modern natural science has grown immeasurably. With the advent of molecular concepts in the structure of matter, probability theory unavoidably came to the fore in both physics and chemistry. From the point of view of molecular physics, every substance consists of an enormous number of small particles in constant motion and in constant interaction. Little is known about the nature of these particles, their interaction, mode of motion and so forth. In general outline, the information about these particles ends with the fact that their numbers are large and that in a homogeneous body their properties are similar. Quite natural, then, that under such conditions the mathematical methods commonly applied to physical theories were helpless. For example, the apparatus of differential equations could not yield serious results in a situation like that. Indeed, neither the structure nor the laws of interaction between the particles of the substance had been sufficiently studied. Application of differential equations in such a situation could only be extremely arbitrary. But even if this difficulty were eliminated, the enormous number of particles represents such a formidable barrier to any study of their motions as to be far beyond the range of the customary equations of mechanics.

What is more, such an approach is methodologically unsatisfactory. Indeed, the problem here is not to study the individual particle motions but to investigate the regularities that arise in assemblies of large numbers of moving and interacting particles. Now the laws that arise from large numbers of participating elements have their own peculiarities and do not reduce to a simple summation of individual motions. Moreover, such regularities are found, within certain limits, to be independent of the individual peculiarities of the participating particles. Quite naturally, fresh mathematical methods of investigation must be found to study these new regularities. What demands are to be made? First, undoubtedly they must take into ac-

count that the phenomenon at hand has to do with large numbers; thus, for these methods, the existence of large numbers of interacting particles should not represent an additional difficulty but rather a simplification for the study of such laws. Further, insufficient knowledge about the nature and structure of the particles and likewise about the nature of their interactions should not limit the efficacy of their application. These demands are best satisfied by the methods of probability theory.

To avoid any misunderstanding, we again stress the following circumstance. When we say that the apparatus of probability theory is best suited to the study of molecular phenomena, we do not in the least wish to say that the philosophical premises for applying the theory of probability in natural science spring from the insufficiency of our knowledge. The basic principle lies in the fact that when studying mass-scale phenomena, a set of *new and peculiar regularities* come to light. When studying phenomena caused by the action of large numbers of molecules, it is not necessary to take into consideration all the properties of every molecule. Indeed, the study of nature requires that *inessential* details be ignored. If all details, all existing relationships, including those that are not essential for the given phenomenon, are considered, then the phenomenon itself is obscured and it becomes more difficult to grasp the subject because of such artificial complications.

We can judge how aptly a phenomenon has been outlined and how successful is the choice of mathematical tools for its study by the agreement between theory and experiment (practice). The development of natural sciences, physics in particular, shows that the apparatus of probability theory has proved exceptionally suited to the study of numerous phenomena of nature.

The above-mentioned relationship between probability theory and the requirements of modern physics best explains the reasons why probability theory during the past few decades has become one of the most rapidly developing branches of mathematics. Fresh theoretical results open up new opportunities for applying methods of probability theory to the natural sciences. Diversified studies of natural phenomena force probability theory to seek new laws generated by the factor of chance. Probability theory does not dissociate itself from the demands of other sciences, but keeps step with the general advance of the natural sciences. This does not, of course, mean that probability theory is only an auxiliary tool in the solution of certain practical problems. Quite the contrary, it must be stressed that during the past three decades, the theory of probability has developed into a harmonious mathematical discipline with its own problems and methods of proof. At the same time it has come to light that the most essential problems of probability theory serve in the solution of diverse problems of natural science and practical affairs.

From the very start we defined the theory of probability as a science concerned with random events. The notion of a random event will be explained in the first chapter. For the time being we confine ourselves to a few remarks. In ordinary parlance a random event is regarded as something extremely rare that runs counter to the established order of things and the law-governed development of events; in the theory of probability, we reject this point of view. In probability theory, random events possess a series of characteristic peculiarities; for one thing, they all occur in mass-scale phenomena. By mass-scale phenomena we have in view such that occur in assemblages of large numbers of entities of equal or nearly equal status and are determined by this mass-scale nature of the phenomenon, depending only in slight measure on the nature of the component entities.

Like the other divisions of mathematics, the theory of probability developed out of the demands of practical affairs; in abstract form it reflects the regularities peculiar to random events of a mass-scale character. Such regularities play an exceedingly important role in physics and in other natural sciences, in military affairs, diversified fields of technology, in economics, and elsewhere. In recent times, in connection with large-scale production, the results of probability theory are not only used in locating defective items already produced but, what is more important, for organizing the very process of production (statistical quality control in production).

As has already been noted, the relationship between probability theory and practical requirements has been the basic reason for the rapid development of probability theory in the past three decades. Many divisions of the theory evolved in response to the problems of practical workers. It is fitting here to recall the remarkable words of the founder of the Russian school of probability theory, P. L. Chebyshev: "The link-up between theory and practice yields the most salutary results, and the practical side is not the only one that benefits; the sciences themselves advance under its influence, for it opens up to them new objects of investigation or fresh aspects of familiar objects... . If the theory gains much from new applications of an old method or from its new developments, then it benefits still more from the discovery of new methods, and in this case too, science finds itself a true guide in practical affairs."

The Concept of Probability

Sec. 1. Certain, Impossible, and Random Events

On the basis of observations and experiment, science arrives at laws that govern the course of the phenomena under study. The most elementary and widespread scheme of the regularities under study is:

1. *In every realization of a set (complex) of conditions \mathfrak{S} there occurs an event A .*

Thus, to illustrate, if water at atmospheric pressure (760 mm) is heated above 100°C (the set of conditions \mathfrak{S}), it is transformed into steam (event A). Or, for any chemical reaction of substances, without exchange with the surrounding medium (the set of conditions \mathfrak{S}), the total quantity of substance (matter) remains unchanged (event A). This assertion is called the law of conservation of matter. The reader will readily be able to point out other laws taken from physics, chemistry, biology and other sciences.

An event that unavoidably occurs for every realization of the set of conditions \mathfrak{S} is called *certain (sure)*. If event A definitely cannot occur upon realization of the set of conditions \mathfrak{S} it is called *impossible*. And if, when the set of conditions \mathfrak{S} is realized, event A may or may not occur, it is called *random*.

From these definitions it is clear that when speaking of certainty, impossibility or randomness of some event, we will always have in view the certainty, impossibility and randomness with respect to some definite set of conditions \mathfrak{S} .

Proposition 1 asserts the certainty of event A upon realization of a set of conditions \mathfrak{S} . Asserting the *impossibility* of some event upon the realization of a given set of conditions does not yield anything essentially new because it readily reduces to assertions of type 1: the impossibility of event A is tantamount to the certainty of the opposite event \bar{A} , which consists in the fact that A does not occur.

The mere assertion of the randomness of an event is of very restricted cognitive interest: it simply amounts to stating that the set of condi-

tions \mathfrak{S} does not reflect the entire collection of reasons necessary and sufficient for the occurrence of A . Such an indication cannot be considered totally devoid of content, for it may serve as a stimulus to further study of the conditions of occurrence of A , but of itself it does not yet yield us any positive knowledge.

However, there is a broad range of phenomena in which, given a repeated realization of the set of conditions \mathfrak{S} , the portion of that range of cases when event A occurs only occasionally deviates to any substantial degree from some average value, which can thus serve as a characteristic indicator of a *mass-scale operation* (multiply repeated set \mathfrak{S} relative to event A).

For such phenomena one can give not only a simple statement of the randomness of event A , but also a quantitative estimation of the possibility of its occurrence. This estimation is expressed by a proposition of the type:

2. *The probability that event A will occur upon realization of a set of conditions \mathfrak{S} is equal to p .*

Regularities of this kind are termed *probabilistic* or *stochastic*. Probabilistic regularities play an important role in diverse fields of science. For instance, there is no way of predicting whether a given radium atom will decay in a given interval of time or not, but it is possible, on the basis of experimental findings, to determine the probability of such decay: an atom of radium decays in a time interval of t years with a probability

$$p = 1 - e^{-0.000436t}$$

Here the set of conditions \mathfrak{S} consists in the fact that we consider a radium atom which for a given number of years is not subjected to any unusual external actions (like bombardment with high-speed particles); otherwise its conditions of existence are inessential: it is of no consequence what the medium is, what temperature it has, and so on. Event A consists in the fact that the atom will decay in the given time of t years.

The idea which now seems to us quite natural, that the probability of a random event A , under known conditions, admits of a quantitative evaluation by means of some number

$$p = P(A)$$

was elaborated in systematic fashion for the first time in the 17th century in the works of Fermat (1601-1665), Pascal (1623-1662), Huygens (1629-1695), and in particular James Bernoulli (1654-1705). Their investigations laid the foundations of the theory of probability. Since that time, the theory of probability has been under development as a mathematical discipline and has become enriched with new important results. Its applicability to the study of actual phenomena of

the most diversified nature continues to find new and brilliant confirmation.

There can be no doubt that the concept of mathematical probability warrants a profound philosophical study. The basic specific philosophical problem advanced by the very existence of probability theory and its successful application to real phenomena consists in the following: *under what conditions is there objective meaning in the quantitative estimate of the probability of a random event A with the aid of a definite number $P(A)$, called the mathematical probability of event A , and what is the objective meaning of this estimate.* A clear understanding of the relationships between the philosophical categories of randomness and necessity is an inevitable prerequisite for a successful analysis of the concept of mathematical probability; but this analysis cannot be complete without answering the question we have posed: under what conditions does chance allow for a quantitative estimate, in the form of a number, of probability.

Every investigator dealing with the application of probability theory to physics, biology, engineering, economic statistics, or any other concrete science, actually proceeds from the conviction that *probabilistic judgements express certain objective properties of the phenomena under study.* To assert that under a certain set of conditions \mathfrak{S} the occurrence of an event A has a probability p is to assert that between the set of conditions \mathfrak{S} and the event A there is a certain perfectly definite, though peculiar (but no less objective for this reason), relationship existing independently of the investigator. The philosophical problem is to elucidate the nature of this relationship. It is only the difficulty of this problem that has made possible the paradoxical circumstance that even among scientists who in general philosophical problems do not take the idealistic stand, one can find attempts to dismiss the problem (instead of solving it in a positive fashion) by asserting that probabilistic judgements have to do only with the state of the investigator (such judgements being regarded as measuring the degree of his confidence that event A will occur, and so forth).

The extensive and diversified experience of applying probability theory in a wide range of fields teaches us that the very problem of a quantitative estimation of the probability of some event has reasonable objective meaning only under certain quite definite conditions.

The definition given above of the randomness of an event A for a set of conditions \mathfrak{S} is purely negative: the event is random if it is not necessary and not impossible. From the fortuitous nature of an event in this purely negative sense it does not in the least follow that it is meaningful to speak of its probability as a certain definite number, even if it is one we do not know. In other words, not only the assertion that "event A has a definite probability $P(A)$ for a set

of conditions \mathfrak{S} ", but also the simple assertion that this probability *exists* is an informative assertion which in each specific case requires substantiation or, when it is taken as a hypothesis, subsequent verification.

For example, a physicist encountering a new radioactive element will from the start assume that for an atom of the element that is left to itself (i.e., not subject to external influences of extraordinarily great intensity) there exists a certain probability of decay during time t , the dependence of which on time is of the form

$$p = 1 - e^{-\alpha t}$$

and will strive to determine the coefficient α that characterizes the rate of decay of the new radioactive element. The question may be posed of the dependence of the probability of decay on external conditions, for example on the intensity of cosmic radiation: here the researcher will proceed from the assumption that to every *sufficiently definite* set of external conditions there corresponds a certain definite value of the coefficient α .

The situation is exactly the same in all other cases of successful practical application of the theory of probability. For this reason, the problem of the philosophical elucidation of the real content of the concept "mathematical probability" may be made hopeless from the very start if a definition is required that may be applied to any event A under any set of conditions \mathfrak{S} .

Sec. 2. Different Approaches to the Definition of Probability

A very large number of different definitions of mathematical probability have been proposed by various authors. We shall not attempt here to examine all the logical niceties of these many definitions. Any scientific definition of such basic concepts as probability is only a subtle logical analysis of a certain store of very simple observations and practical procedures that have justified themselves by long and successful employment. Interest in a logically irreproachable "substantiation" of probability theory arose later, historically speaking, than the ability to determine the probability of various events, to perform calculations with these probabilities and also to utilize the results of the calculations in practical affairs and in scientific research. For this reason, in most attempts at a scientific definition of the general concept of probability it is easy to perceive various aspects of the concrete cognitive process which in each specific case leads to an actual determination of the probability of a given event, whether this is the probability of getting a six in four throws of a die or the probability of radioactive decay or of hitting a target. Some definitions start from inessential, subsidiary aspects of real processes; these

are totally fruitless. Others advance some one aspect of the matter or certain modes of actually finding the probability that are not applicable in all cases. These definitions merit closer examination despite their one-sided nature.

From the viewpoint thus delineated, most definitions of mathematical probability may be divided into three groups:

1. Definitions of mathematical probability as a quantitative measure of the "degree of certainty" of the observer.

2. Definitions that reduce the concept of probability to that of "equal possibility" as being the most primitive concept (the so-called "*classical*" definition of probability).

3. Definitions that proceed from the "frequency" of occurrence of an event in a large number of trials (the "*statistical*" definition).

Sections 4 and 7 are devoted to the second and third groups. The definitions of the first group will be critically examined at the end of this section. If mathematical probability is a quantitative measure of the degree of certainty of the investigator, then the theory of probability will reduce to something in the nature of a division of psychology. Ultimately, a consistent development of such a purely subjectivistic conception of probability would unavoidably lead to subjective idealism. Indeed, if it is assumed that the evaluation of probability is related only to the state of the investigator, then all conclusions from probabilistic judgements (judgements of type 2) are stripped of the objective content that is independent of the investigator. Yet type 2 probabilistic judgements are used as a basis for many positive conclusions which in no way differ in significance from conclusions obtained without appeal to probability. For example, physics derives all the "macroscopic" properties of gases from suppositions concerning the nature of the probabilities of behaviour of the individual molecules. If we attribute objective value (that is, value independent of the investigator), then in the initial probabilistic hypotheses concerning the course of "macroscopic" molecular processes it is necessary to perceive something more important than mere registration of our psychological states that arise when meditating about molecular motions.

For those who take the stand of reality existing independently of us and of the fundamental knowability of the external world, and who also reckon with the fact that probabilistic judgements are successfully employed in learning about the external world, it must be absolutely clear that the purely subjective definition of mathematical probability is quite untenable. This might suffice to complete the discussion of definitions of the first group if they did not find support in the original everyday meaning of the word "probability". The point is that in ordinary usage the expressions "probably", "very probably", "it is highly improbable", and so forth do indeed express simply the attitude of the speaker to the truth or falsity of some

single judgement. We must therefore put stress on a circumstance that we have not dwelt on as yet. When in Section I we immediately centred attention on type 2 probabilistic regularities by opposing them to the strict causal regularities of type 1, we acted in full accord with practical applications of the concept of mathematical probability, but from the very start we somewhat digressed from the ordinary "prescientific" meaning of the word "probability": whereas in all real scientific applications of probability theory, "probability" is the probability of the occurrence of some event A , provided that a certain set of conditions \mathfrak{S} , which is *fundamentally reproducible an infinite number of times*, has been realized (it is only in such a setting that the statement

$$p = P(A)$$

expresses a certain regularity with objective meaning), in ordinary parlance the customary thing is to speak of a greater or lesser probability of some *quite definite* judgement. For example, relative to the judgements:

(a) every even natural number greater than two may be represented in the form of a sum of two prime numbers ($4=2+2$, $6=3+3$, $8=5+3$, etc.);

(b) snow will fall in Moscow on May 7, 1976; the following may be said: concerning judgement (a) we do not yet have full knowledge, but many believe it to be extremely likely; one must expect to get the exact answer to judgement (b) only on 7 May 1976. However, since snow very rarely falls in Moscow in May, judgement (b) should for the present be considered highly unlikely.

Indeed, we only attribute a subjective meaning to such statements relative to the probability of isolated facts or, in general, specific judgements (even though of a general nature): they reflect only the attitude of the speaker to the given question. And it is true that when speaking about the greater or lesser probability of a specific judgement, we ordinarily do not in the least desire to doubt that the law of the excluded middle is applicable. For instance, no one doubts that each of the propositions (a) and (b) is actually *true* or *false*. Even if such doubts were expressed by the so-called intuitionists concerning judgements (a), at any rate, to the ordinary mind, the possibility of speaking of a greater or lesser probability of this proposition is in no way related to doubts as to whether the law of the excluded middle is applicable or not. If proposition (a) is ever proved or refuted, all preliminary estimates of its probability made at the present time become meaningless. In exactly the same way, when 7 May 1976 comes, it will be easy to see whether (b) is true or not; if snow falls on that day, there will be no sense in giving the view that this event is improbable.

A complete investigation of the extreme diversity of psychic states of *doubt* intermediate between a categorical admission and categorical rejection of an isolated judgement, no matter how interesting that may be for psychology, would only lead us far astray from our basic problem of elucidating the meaning of probabilistic regularities of objective scientific value.

Sec. 3. The Sample Space

In the preceding section we saw that the definition of mathematical probability as a quantitative measure of the "degree of certainty" of the investigator does not capture the content of the notion of probability. We therefore return to the question of where objective probabilistic regularities come from. The classical and statistical definitions of probability claim to yield simple and direct answers to this question. We shall see later on that both these definitions reflect essential aspects of the actual content of the notion of probability, though each one taken separately is insufficient. A full understanding of the nature of probability demands their synthesis. In the next few sections we shall deal exclusively with the classical definition of probability that proceeds from the idea of *equal likelihood* as an objective property of diverse possible versions of the course of phenomena based on their actual symmetry. Henceforward we shall have to do only with such a conception of equal likelihood. The definition of probability in terms of "equal likelihood" as understood in a purely subjective sense of the identical "likelihood" for the investigator is a variant of the definitions of probability expressed in terms of "degree of certainty" of the observer which we have already dismissed from our consideration.

Before passing to the classical definition of the notion of probability we shall make some preliminary remarks. We will consider as fixed a set of conditions \mathfrak{S} and will examine a certain family S of events A, B, C, \dots^* , each of which must *occur or not occur* upon realization of the set \mathfrak{S} . Certain relationships may obtain between the events of the family S . Since they will be constantly under study later on, let us look into them at the outset.

(1) If for every realization of a set of conditions \mathfrak{S} under which an event A occurs, there also occurs an event B , then we say that A *implies* B and we denote this by the symbol

$$A \subset B$$

or

$$B \supset A$$

which means that " B is implied by A ".

* Events will always be designated by capital letters A, B, C, D , etc.

(2) If A implies B and at the same time B implies A , that is, if in each realization of the set of conditions \mathfrak{S} events A and B both occur or both fail to occur, we shall say that events A and B are *equivalent*; this will be denoted by the symbol $A=B$.

We note that in all considerations of probability theory, equivalent events can replace one another. For this reason, we will agree in the future to consider any two equivalent events as simply identical.

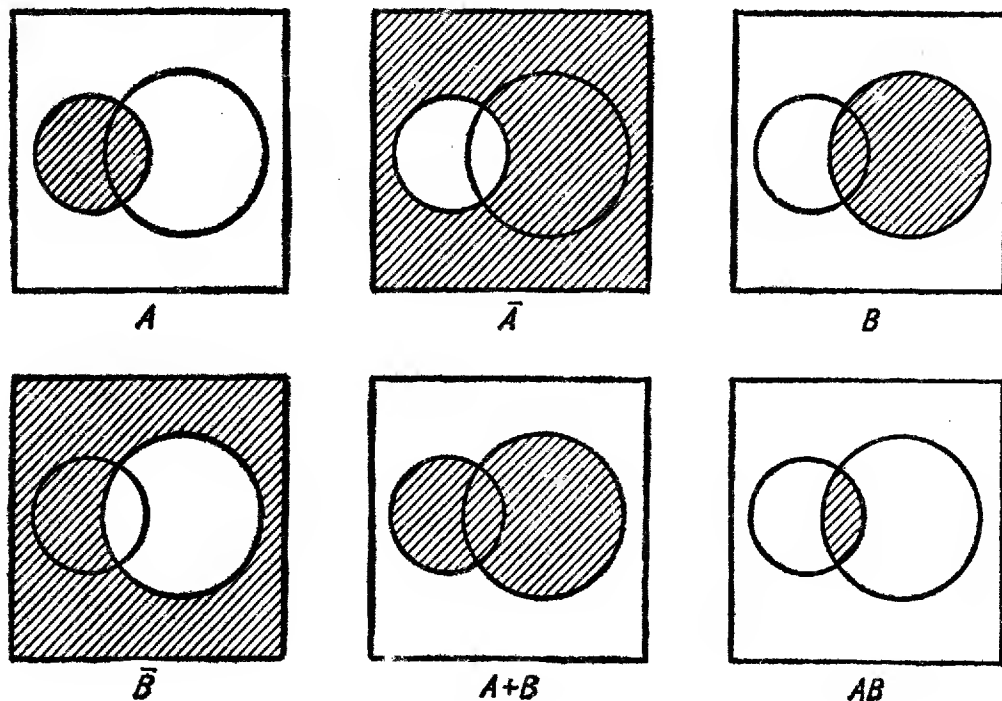


Fig. 1

(3) An event which consists in the occurrence of both events A and B will be called the *product* of A and B and will be designated AB (or $A \cap B$).

(4) An event consisting in the occurrence of at least one of the events A and B will be called the *sum* of A and B and will be designated $A+B$ (or $A \cup B$).

(5) An event which consists in the fact that event A occurs but event B does not will be called the *difference* of the events A and B and will be designated $A - B$.

We shall illustrate the newly introduced concepts with simple examples. The first is what is known as the *Venn diagram*.

Let there be a set of conditions \mathfrak{S} such that a point is chosen at random within a square, as depicted in Fig. 1. Denote by A the event that "the chosen point lies inside the circle on the left", and by B the event that "the chosen point lies inside the circle on the right". Then events A , \bar{A} , B , \bar{B} , $A+B$, AB consist in the chosen point being located in the regions shaded in the appropriate squares of Fig. 1.

Let us take another illustration. Suppose that the set of conditions \mathfrak{S} consists in the fact that a die* is tossed once. We denote by A a roll of six, by B a roll of three, by C a roll of some even number of points, by D the roll of a sum that is a multiple of three. Then events A , B , C and D are connected by the following relations:

$$A \subset C, A \subset D, B \subset D \\ A + B = D, CD = A$$

The definition of a sum and product of two events is generalized to any number of events:

$$A + B + \dots + N$$

This signifies an event that consists in the occurrence of at least one of the events A , B , \dots , N , and

$$AB \dots N$$

signifies an event consisting in the occurrence of all events A , B , \dots , N .

(6) An event is called *certain* if it of necessity must occur (upon realization of a set of conditions \mathfrak{S}). For example, in throwing two dice the total sum of points must certainly be at least two.

An event is called *impossible* if it definitely cannot occur (no matter what the realization of the set of conditions \mathfrak{S}). For instance, the throwing of two dice can never yield a sum of thirteen points.

All certain events are obviously equivalent. It is therefore justifiable to denote all certain events by a single letter. For this purpose we shall use the letter U . All impossible events are likewise equivalent. We denote an impossible event by V .

(7) Two events A and \bar{A} are termed *contrary* if for them the two following relationships hold simultaneously:

$$A + \bar{A} = U, A\bar{A} = V$$

For example, if throwing a die C signifies a roll with an even sum, then

$$U - C = \bar{C}$$

is an event consisting in a roll with an odd sum of points.

(8) Two events A and B are called *mutually exclusive* if their joint occurrence is impossible, that is, if

$$AB = V$$

If

$$A = B_1 + B_2 + \dots + B_n$$

* A die is a cube the faces of which are numbered 1, 2, 3, 4, 5 and 6.

and events B_i are mutually exclusive in pairs, i.e.,

$$B_i B_j = V \text{ for } i \neq j$$

then we say that event A is *decomposable into the particular events* B_1, B_2, \dots, B_n . To illustrate, when tossing a die, event C , which consists in a roll of an even sum of points, is decomposable into the special events E_2, E_4, E_6 , which are, respectively, rolls of 2, 4 and 6.

The events B_1, B_2, \dots, B_n form a *complete group of events* if at least one of them must definitely occur (for each realization of the set \mathfrak{S}); that is if

$$B_1 + B_2 + \dots + B_n = U$$

Of special significance to us in the sequel will be *complete groups of pairwise mutually exclusive events*. Such, for example in a single toss of one die, is the family of events

$$E_1, E_2, E_3, E_4, E_5, E_6$$

which consists in rolls of 1, 2, 3, 4, 5, and 6 points, respectively.

(9) Every problem in probability theory involves a certain set of conditions \mathfrak{S} and a certain family S of events that occur or do not occur upon every realization of the set of conditions \mathfrak{S} . It is advisable to make the following assumptions relative to such a system:

(a) *if the family S includes events A and B , it also includes events $AB, A+B, A-B$;*

(b) *the family S contains a certain and an impossible event.*

A family of events that satisfies these assumptions is called a *field of events*.

In the examples that we have used for illustrations, it was always possible to isolate events that could not be decomposed into simpler ones: a certain face turning up in the throw of a die; landing on a definite point in the square when considering Venn's diagram. Let us agree to call such indecomposable events *simple (elementary) events*.

In constructing a mathematical theory of probability, our intuitive conceptions demand greater formalization than heretofore. In the modern expositions of probability theory, the starting point is a set of simple events, or, as it is generally termed, a *space of simple events* (a *sample space*). The nature of the elements of this space is not specified beforehand inasmuch as it is important to have a sufficiently broad choice so as to embrace all possible cases. For instance, the elements of the space may be points of Euclidean space, the functions of one or several variables, and so forth. The sets of points of a sample space form random events. An event which consists of all the points of a sample space is called a *certain (sure) event*. Everything that we have said about the relations between random events in this section also holds true for a formal construction of the theory. We shall return to this system of exposition somewhat later, in Se-

ction 8. In the next section we will confine ourselves to sample spaces consisting of a finite number of elements.

Here we shall confine ourselves to stating the following laws that hold for random events:

Commutative law: $A+B=B+A$, $AB=BA$.

Associative law: $A+(B+C)=(A+B)+C$, $A(BC)=(AB)C$.

Distributive law: $A(B+C)=AB+AC$, $A+(BC)=(A+B)(A+C)$.

Idempotency law: $A+A=A$, $AA=A$.

We leave the proof of these laws to the reader. For anyone acquainted with elementary set theory there will be no difficulty.

Sec. 4. The Classical Definition of Probability

The classical definition of probability reduces the concept of probability to the notion of equal probability (equal likelihood) of events, which is considered basic and is not subject to a formal definition. To illustrate, when tossing a die that has the exact shape of a cube and is made of completely homogeneous material, equally probable events are rolls of any specific sum of points (1, 2, 3, 4, 5, 6) marked on the faces of the cube, since by virtue of symmetry no face has an objective advantage over any of the others.

In the general case, we consider some group G consisting of n mutually exclusive equally probable events (we call them *simple events*):

$$E_1, E_2, \dots, E_n$$

We now form a family S consisting of the impossible event V , all events E_k of group G and all events A that may be decomposed into special cases belonging to the group G .

For example, if G consists of three events E_1, E_2 and E_3 , then system S includes the events* $V, E_1, E_2, E_3, E_1+E_2, E_2+E_3, E_1+E_3, U=E_1+E_2+E_3$.

It is readily established that the family S is a field of events. Indeed, it is obvious that the sum, difference and product of events of S are included in S ; the impossible event V belongs to S by definition, and the certain event U belongs to S , since it is represented in the form

$$U=E_1+E_2+\dots+E_n$$

The classical definition of probability is given for events of family S and may be formulated as follows:

* These eight events exhaust the family S provided we do not distinguish (as we agreed to do at the end of Sec. 2) equivalent events. It will readily be shown that in the general case of a group G of n events, the family S consists of 2^n events.

If event A is decomposable into m special cases belonging to a complete group of n mutually exclusive and equally probable events, the probability $P(A)$ of event A is

$$P(A) = \frac{m}{n}$$

For example, in a single toss of a die, the complete group of mutually exclusive and equally probable events comprises the events

$$E_1, E_2, E_3, E_4, E_5, E_6$$

which consist of rolls of 1, 2, 3, 4, 5, 6 points, respectively. Event

$$C = E_2 + E_4 + E_6$$

which consists of an even number of points is made up of three special cases that are components of the complete group of mutually exclusive and equally probable events. Therefore, the probability of event C is

$$P(C) = \frac{3}{6} = \frac{1}{2}$$

It is also obvious that by virtue of the accepted definition

$$P(E_i) = \frac{1}{6}, \quad 1 \leq i \leq 6,$$

$$P(E_1 + E_2) = \frac{2}{6} = \frac{1}{3}$$

and so on.

The theory of probability makes wide use of the following terminology which we will need later on. Suppose that in order to find out whether an event A (say rolls of multiples of three) occurs or does not occur it is necessary to make a *trial* (that is, realize a set of conditions \odot) that would give us the answer (in our case, throwing a die). The complete group of mutually exclusive and equally probable events that can occur in such a trial is called the complete group of *possible outcomes* of the trial. The possible outcomes of a trial into which event A is decomposed are called the outcomes of the trial that are favourable to A . Employing this terminology, we can say that the *probability $P(A)$ of event A is equal to the number of possible trial outcomes favourable to A divided by the number of all possible outcomes*.

Such a definition quite naturally presumes that the separate possible outcomes are equally probable.

Now let us consider the throwing of two dice. If the dice are true, a roll of each of the 36 possible combinations of numbers on both dice may be considered equally probable. Say, the probability of rolling 12 is equal to $1/36$. A roll of 11 is possible in two ways: 5 on

the first die, 6 on the second, or 6 on the first and 5 on the second. Therefore the probability of a roll of eleven is $2/36=1/18$. The reader will find it easy to verify that the probability of any specific roll (any sum) is given by the following table:

TABLE 1

Sum	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

In accord with this definition, to every event A that belongs to the thus constructed field of events S there is ascribed a very definite probability

$$P(A) = \frac{m}{n}$$

where m is the number of those events E_i of the initial group G that are special cases of event A . Thus, the probability $P(A)$ may be regarded as a *function of the event A defined on the field of events S* .

This function possesses the following properties:

1. For every event A of field S

$$P(A) \geq 0$$

2. For the certain event U

$$P(U) = 1$$

3. If event A is decomposable into special cases B and C and all three events A , B and C belong to the field S , then

$$P(A) = P(B) + P(C)$$

This property is called the *theorem of addition of probabilities*.

Property 1 is obvious since the fraction $\frac{m}{n}$ cannot be negative. Property 2 is no less obvious since all the n possible outcomes of a trial are favourable to the certain event U , and therefore

$$P(U) = \frac{n}{n} = 1$$

We shall prove Property 3. Let event B be favoured by m' events, and event C by m'' events E_i of group G . Since by assumption the events B and C are mutually exclusive, events E_i , which favour one of them, are different from events E_j that favour the

other. There is thus a total of $m' + m''$ events E_i favourable to the occurrence of one of the events B and C , i.e., favourable to the event $B + C = A$. Hence,

$$P(A) = \frac{m' + m''}{n} = \frac{m'}{n} + \frac{m''}{n} = P(B) + P(C)$$

which is what we had to prove.

We restrict ourselves here to a few more properties of probability.

4. *The probability of event \bar{A} , which is the opposite of event A , is*

$$P(\bar{A}) = 1 - P(A)$$

Indeed, since $A + \bar{A} = U$, we have, from the already proved Property 2,

$$P(A + \bar{A}) = 1$$

and since events A and \bar{A} are mutually exclusive, by Property 3

$$P(A + \bar{A}) = P(A) + P(\bar{A})$$

The last two equations prove our proposition.

5. *The probability of the impossible event is zero.*

Indeed, events U and V are mutually exclusive, therefore

$$P(U) + P(V) = P(U)$$

whence it follows that

$$P(V) = 0$$

6. *If event A implies event B , then*

$$P(A) \leq P(B)$$

Indeed, event B may be represented as the sum of two events A and $\bar{A}B$. From this, by virtue of Properties 3 and 1, we have

$$P(B) = P(A + \bar{A}B) = P(A) + P(\bar{A}B) \geq P(A)$$

7. *The probability of any event lies between zero and unity.*

From the fact that for any event A the relations

$$V \subset A + V = A = AU \subset U$$

hold, there follow, by virtue of the preceding property, the inequalities

$$0 = P(V) \leq P(A) \leq P(U) = 1$$

Sec. 5. The Classical Definition of Probability. Examples

We now consider some cases in calculating the probabilities of events using the classical definition of probability. The examples

are strictly illustrative in character and do not claim to exhaust all basic methods of calculating probabilities.

Example 1. From a deck of 36 cards draw three at random. Find the probability that there will be exactly one ace among them.

Solution. The complete group of equally probable and mutually exclusive events in our problem consists of all possible combinations of three cards; their number is C_{36}^3 . The number of favourable events may be computed as follows. One ace may be drawn in C_4^1 different ways, while the two other cards (non-aces) may be drawn in C_{32}^2 different ways. Since for each definite ace the two remaining cards may be chosen in C_{32}^2 ways, there will be a total of $C_4^1 \cdot C_{32}^2$ favourable cases. Thus, the desired probability will be

$$p = \frac{C_4^1 \cdot C_{32}^2}{C_{36}^3} = \frac{\frac{4}{1} \cdot \frac{32 \cdot 31}{1 \cdot 2}}{\frac{36 \cdot 35 \cdot 34}{1 \cdot 2 \cdot 3}} = \frac{31 \cdot 16}{35 \cdot 3 \cdot 17} = \frac{496}{1785} \approx 0.2778$$

which is slightly more than 0.25.

Example 2. From a deck of 36 cards draw three at random. Find the probability that there will be at least one ace.

Solution. Denote the event we are interested in by A : it may be represented in the form of a sum of the three following mutually exclusive events: A_1 the occurrence of one ace, A_2 the occurrence of two aces, A_3 the occurrence of three aces.

By reasoning similar to that given in the solution of the previous problem it is easy to establish that the number of cases favourable to the

event A_1 is $C_4^1 \cdot C_{32}^2$

event A_2 is $C_4^2 \cdot C_{32}^1$

event A_3 is $C_4^3 \cdot C_{32}^0$

Since the number of all possible cases is C_{36}^3 , we have

$$P(A_1) = \frac{C_4^1 \cdot C_{32}^2}{C_{36}^3} = \frac{16 \cdot 31}{3 \cdot 35 \cdot 17} \approx 0.2778$$

$$P(A_2) = \frac{C_4^2 \cdot C_{32}^1}{C_{36}^3} = \frac{3 \cdot 16}{3 \cdot 35 \cdot 17} \approx 0.0269$$

$$P(A_3) = \frac{C_4^3 \cdot C_{32}^0}{C_{36}^3} = \frac{1}{3 \cdot 35 \cdot 17} \approx 0.0006$$

By virtue of the addition theorem,

$$P(A) = P(A_1) + P(A_2) + P(A_3) = \frac{109}{3 \cdot 119} \approx 0.3053$$

This example may be solved in yet another way. Event \bar{A} , the opposite of A , consists in the fact that there will not be a single ace among the drawn cards. Obviously, three non-aces can be drawn from a deck of 36 cards in C_{32}^3 different ways and, hence,

$$P(\bar{A}) = \frac{C_{32}^3}{C_{36}^3} = \frac{32 \cdot 31 \cdot 30}{36 \cdot 35 \cdot 34} = \frac{31 \cdot 8}{3 \cdot 17 \cdot 7} \approx 0.6947$$

The desired probability is

$$P(A) = 1 - P(\bar{A}) \approx 0.3053$$

Note. In both instances the expression “at random” means that all possible combinations of three cards are equally probable.

Example 3. A deck of 36 cards is divided at random into two equal parts. What is the probability that both parts will have an equal number of red and black cards?

The expression “at random” means that all possible divisions of the deck into two equal parts are equally likely.

Solution. We have to find the probability that of 18 cards drawn at random from the deck 9 will be red and 9 black.

The total number of different ways to draw 18 cards from 36 is C_{36}^{18} . The favourable ways are those in which there will be 9 cards drawn from 18 red cards and 9 from 18 black cards. Nine red cards may be drawn in C_{18}^9 different ways and 9 black cards also in C_{18}^9 different ways. Since in drawing 9 definite red cards, 9 black ones may be drawn in C_{18}^9 different ways, the total number of favourable ways is equal to $C_{18}^9 \cdot C_{18}^9$. And consequently the sought-for probability is

$$p = \frac{C_{18}^9 \cdot C_{18}^9}{C_{36}^{18}} = \frac{(18!)^4}{36! (9!)^4}$$

In order to get an idea of the magnitude of this probability without performing arduous computations, we make use of Stirling's formula which yields the following asymptotic relation:

$$n! \sim \sqrt{2\pi n} n^n e^{-n}$$

We thus have

$$18! \approx 18^{18} e^{-18} \sqrt{2\pi \cdot 18}$$

$$9! \approx 9^9 e^{-9} \sqrt{2\pi \cdot 9}$$

$$36! \approx 36^{36} e^{-36} \sqrt{2\pi \cdot 36}$$

and consequently

$$p \approx \frac{(\sqrt{2\pi \cdot 18} \cdot 18^{18} \cdot e^{-18})^4}{\sqrt{2\pi \cdot 36} \cdot 36^{36} \cdot e^{-36} (\sqrt{2\pi \cdot 9} \cdot 9^9 \cdot e^{-9})^4}$$

After some simplifications we find that

$$p \approx \frac{2}{\sqrt{18\pi}} \approx \frac{4}{15} \approx 0.26$$

Example 4. There are n particles, each of which can occupy each of N ($N > n$) cells with the same probability $\frac{1}{N}$. Find the probability that: (1) there will be one particle in each of n definite cells, (2) there will be one particle in each of n arbitrary cells.

Solution. This problem plays an important role in modern statistical physics; and depending on how the complete group of equally probable events is formed we have one or another physical statistics: Boltzmann, Bose-Einstein, Fermi-Dirac.

In Boltzmann statistics, any thinkable distributions that differ not only as to number but also as to the individuality of the particles are equally probable: each cell can accommodate any number of particles from 0 to n .

The total number of possible distributions may be computed in the following way: each particle may be located in each of the N cells; hence, n particles may be distributed in the cells in N^n different ways.

In the first question, the number of favourable cases will obviously be $n!$ and, consequently, the probability that one particle will fall in n definite cells is

$$p_1 = \frac{n!}{N^n}$$

In the second question, the number of favourable cases is C_N^n times greater and hence the probability that there will be one particle in n arbitrary cells is equal to

$$p_2 = \frac{C_N^n \cdot n!}{N^n} = \frac{N!}{N^n (N-n)!}$$

In Bose-Einstein statistics identical cases are those in which the particles change places among the cells (the only important thing is how many particles there are in a cell but not the individuality of the particles), and the complete group of equally probable events consists of all possible distributions of n particles in N cells, one distribution being the whole class of Boltzmann distributions, which differ not in numbers of particles in specific cells but only in the identity of the particles themselves. To get a clear-cut idea of the difference between Boltzmann statistics and Bose-Einstein statistics consider a special case: $N=4$, $n=2$. All possible distributions in this case may be written in the form of a table (see below) in which a and b are the names of the particles. In Boltzmann statistics all

16 possibilities are different equally probable events, while in Bose-Einstein statistics the cases 5 and 11, 6 and 12, 7 and 13, 8 and 14, 9 and 15, 10 and 16 are identical pairs and we have a group of 10 equally probable events.

Now compute the total number of equally probable cases in Bose-Einstein statistics. We note that all possible distributions of particles in cells may be obtained as follows: arrange the cells in a straight line, one after the other, then arrange the particles on the same straight line one next to the other. Now consider all possible permutations of particles and partitions between the cells. It is then easy to see that all possible fillings of cells differing both as to order of particles in the cells and as to order of partitions will be taken into account.

The number of these permutations is $(N+n-1)!$. They include identical permutations, in which each distribution among the cells is counted $(N-1)!$ times, since we distinguished the partitions between the cells and also counted each distribution in the cells $n!$ times, because we took into account not only the number of particles in a cell but also the kind of particles and their order. We thus counted every cell distribution $n!(N-1)!$ times, whence the number of different (in the Bose-Einstein sense) distributions of particles in the cells is

$$\frac{(n+N-1)!}{n!(N-1)!}$$

Thus the number of equally probable events in a complete

TABLE 2

Cases	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Cells	<i>ab</i>				<i>a</i>	<i>a</i>	<i>a</i>				<i>b</i>	<i>b</i>	<i>b</i>			
		<i>ab</i>			<i>b</i>			<i>a</i>	<i>a</i>		<i>a</i>			<i>b</i>	<i>b</i>	
			<i>ab</i>			<i>b</i>		<i>b</i>		<i>a</i>		<i>a</i>		<i>a</i>		<i>b</i>
				<i>ab</i>			<i>b</i>		<i>b</i>	<i>b</i>			<i>a</i>		<i>a</i>	<i>a</i>

group of events has been found. It is now easy for us to answer the questions of our problem. In Bose-Einstein statistics the pro-

babilities p_1 and p_2 are

$$p_1 = \frac{1}{\frac{(n+N-1)!}{n!(n-1)!}} = \frac{n!(N-1)!}{(n+N-1)!}$$

$$p_2 = \frac{C_N^n}{\frac{(n+N-1)!}{n!(N-1)!}} = \frac{N!(N-1)!}{(N-n)!(N+n-1)!}$$

Finally, we consider Fermi-Dirac statistics. According to this statistics, each cell accommodates either one particle or none: the individuality of the particle is ignored.

The total number of distinct particle distributions in cells in Fermi-Dirac statistics is computed with ease: the first particle may be distributed in N different ways, the second only in $N-1$, the third in $(N-2)$ and, finally, the n th in $(N-n+1)$ different ways. Here, the different ways are taken as the modes of distribution that differ only in the permutation of particles in the cells. To eliminate particle individuality we must divide the number thus obtained by $n!$.

Then n particles will be arranged in N cells in

$$\frac{1}{n!} \cdot N(N-1) \dots (N-n+1) = \frac{N!}{(N-1)!n!}$$

distinct equally probable ways.

It is easy to figure out that in Fermi-Dirac statistics the sought-for probabilities are

$$p_1 = \frac{(N-n)!n!}{N!}$$

$$p_2 = 1$$

The foregoing example shows how important it is to define exactly what events are considered equally probable in a problem.

Example 5. At a theatre ticket office is a queue of $2n$ persons, n have only five-ruble bills and the remaining n have only 10-ruble bills. The ticket seller has no change to begin with and each customer takes only one 5-ruble ticket. What is the probability that not a single customer will expect change?

Solution. All possible arrangements of customers are equally probable. We employ the following geometric procedure: regard an xy -plane and assume that the customers are arranged along points of the x -axis with coordinates $1, 2, \dots, 2n$ as they stand in the queue. The ticket office is located at the origin. To each person with a 10-ruble bill we ascribe an ordinate of 1 and to each with a 5-ruble bill an ordinate of -1 . Add from left to right the ordinates thus defined at integ-

ral points and plot at each of them the sum obtained (Fig. 2). It will readily be seen that at the point with abscissa $2n$ the sum is 0 (there will be n terms equal to $+1$, and n terms equal to -1). Now connect with straight lines the adjacent points thus obtained and also connect the origin with the leftmost point. The broken line thus obtained will be called the trajectory.

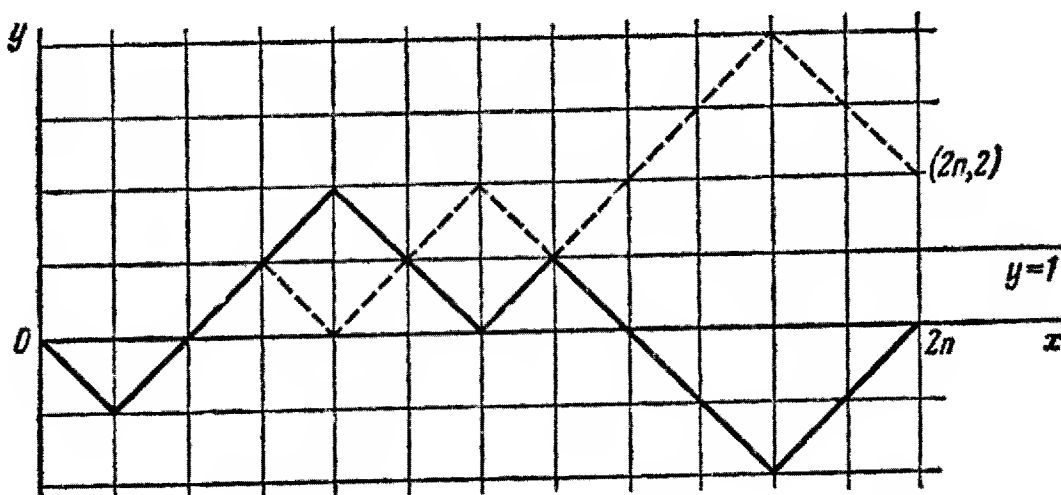


Fig. 2

The total number of distinct possible trajectories, as will be readily seen, is C_{2n}^n (it is equal to the number of all possible orderings of n ascents among $2n$ descents and ascents). The favourable trajectories in our case will be those that do not rise above the abscissa axis (otherwise at least one customer will approach the ticket seller when there is no change available).

Calculate the total number of trajectories which at least once reach or cross the line-segment $y=1$. For this purpose we construct a new (fictitious) trajectory as follows: prior to first contact with the line $y=1$ the new trajectory will coincide with the original one; from the point of contact the new trajectory is a mirror image of the old trajectory relative to the line $y=1$ (see the dashed broken line in Fig. 2). It is easy to see that a trajectory is defined only for trajectories that have at least once reached the line $y=1$, while for the remaining trajectories (i. e., those favourable to our event) it coincides with the original one. Further, the new trajectory begins at point $(0, 0)$ and ends at point $(2n, 2)$. It thus has two more single ascents than descents. Hence the total number of new trajectories is C_{2n}^{n+1} (the number of orderings of $n+1$ ascents among $2n$ ascents and descents). Thus the number of favourable cases is $C_{2n}^n - C_{2n}^{n+1}$ and the sought-for probability is

$$p = \frac{C_{2n}^n - C_{2n}^{n+1}}{C_{2n}^n} = 1 - \frac{n}{n+1} = \frac{1}{n+1}$$

Sec. 6. Geometrical Probability

From the very beginning of the development of probability theory it was noticed that the "classical" definition of probability based on the consideration of a finite group of equally probable events was insufficient. Even at that time special examples led to a certain modification of the definition and to the construction of a concept of probability for cases in which even an infinite set of outcomes is conceivable. As before, the concept of an "equal probability" of some events played the basic role.

The general problem that was posed and that led to an extension of the notion of probability may be formulated as follows.

On a plane, let there be a certain region G and in it another region g with a rectifiable boundary. A point is thrown at random onto G and we wish to find the probability that the point will fall in region g . Here, the expression "a point is thrown at random onto region G " is given the following meaning: the thrown point can fall in any point of G , the probability of falling in any portion of region G is proportional to the measure of this part (length, area, etc.) and is independent of its position and shape.

Thus, by definition, the probability that a point thrown randomly onto G will fall in g is equal to

$$p = \frac{\text{mes } g}{\text{mes } G}$$

Let us consider some examples.

Example 1. The Encounter Problem. Two persons A and B have agreed to meet at a definite spot between 12 and one o'clock. The first one to come waits for 20 minutes and then leaves. What is the probability of a meeting between A and B if the arrival of each during the indicated hour can occur at random and the times of arrival are independent*.

Solution. Denote the times of arrival of A by x and of B by y . For the meeting to take place, it is necessary and sufficient that

$$|x - y| \leq 20$$

We depict x and y as Cartesian coordinates in the plane; for the scale unit we take one minute. All possible outcomes will be described as points of a square with side 60; favourable outcomes will lie in the shaded region (Fig. 3).

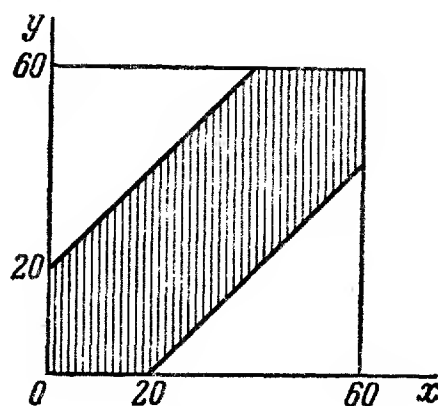


Fig. 3

* That is, the time of arrival of one person does not affect the arrival time of the other.

The desired probability * is equal to the ratio of the area of the shaded figure to the area of the whole square:

$$p = \frac{60^2 - 40^2}{60^2} = \frac{5}{9}$$

Some research engineers have applied the encounter problem to solving problems in the organization of production. A workman is in charge of a number of machines of one type, each of which can at random moments demand his attention. It may happen that when he is busy with one machine, others will be needing his attention. It is required to find the probability of this event; that is, in other words, to find how long on the average the machine waits for the workman (or the time the machine is idle). However, it may be noted that the scheme of the encounter problem is not well suited for a solution of this production problem because there is no agreed-upon time during which the machines definitely require the attention of the workman, and the times the workman spends at any one machine are not constant. Aside from this basic reason, we can point to the complexity of calculations in the encounter problem for the case of a large number of persons (machines). This case often comes up (in the textile industry, for instance, some weavers operate up to 280 looms).

The theory of geometrical probability has repeatedly been criticized for arbitrariness in determining the probabilities of events. Many authors have become convinced that for an infinite number of outcomes the probability cannot be determined objectively, that is, independently of the mode of computation. A particularly brilliant proponent of this scepticism is the French mathematician of the 19th century Joseph Bertrand. In his course of probability theory he cited a number of problems on geometrical probability in which the result depended on the mode of solution. The following is an illustration.

Example 2. Bertrand's Paradox. A chord is chosen at random in a circle. What is the probability that its length will exceed the length of the side of the equilateral triangle inscribed in the circle?

Solution 1. For reasons of symmetry we can specify the direction of the chord beforehand. Draw a diameter perpendicular to this direction. It is obvious that only chords that intersect the diameter in the interval between one fourth and three fourths its length will

* In Sec. 9 we will see that by virtue of the independent arrival times of A and B the probability that A will arrive in the interval from x to $x+h$, and B within the interval between y and $y+s$ is equal to $\frac{h}{60} \cdot \frac{s}{60}$, that is, it is proportional to the area of a rectangle with sides h and s .

exceed the side of the regular triangle. The desired probability is thus $\frac{1}{2}$.

Solution 2. Reasoning from symmetry, it is possible to fix one of the ends of the chord on the circle in advance. The tangent to the circle at this point and two sides of the regular triangle with vertex in this point form three 60° angles. Only chords falling in the middle angle are favourable to the conditions of the problem. For this mode of calculation, the sought-for probability will come out to $1/3$.

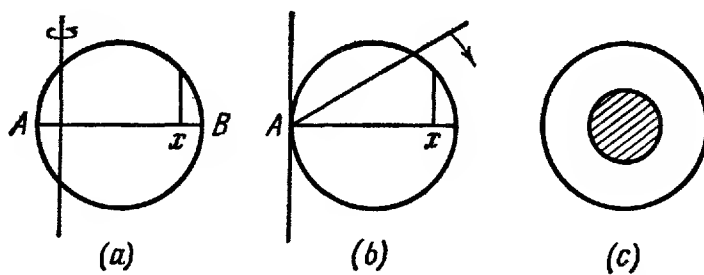


Fig. 4

Solution 3. To define the position of the chord it is sufficient to specify its midpoint. So that the chord will satisfy the condition of the problem, it is necessary that its midpoint lie inside a circle concentric with the given one but with one-half the radius. The area of this circle is equal to one-fourth the area of the given one. Thus, the sought-for probability is $\frac{1}{4}$.

We now have to find out wherein lies the nonuniqueness of the solution of our problem. Is the reason a fundamental impossibility to determine the probability for cases of an infinite number of possible outcomes or is it because some of our starting premises were impermissible?

It is easy to see that solutions of three different problems are given as the solution of one and the same problem due to the fact that the conditions of the problem do not define the notion of drawing a chord at random.

Indeed, in the first solution, a circular cylindrical rod (Fig. 4a) is made to roll along one of the diameters. The set of all possible stopping points of this rod forms the set of points of a segment AB of length equal to the diameter. Equiprobable events are those which consist in a stop occurring in an interval of length h , no matter where this segment is located on the diameter.

In the second solution, a rod fixed on a hinge situated on one of the points of the circle is made to oscillate no more than 180° (Fig. 4b). It is assumed here that a stop of the rod inside an arc of length h

of the circle depends solely on the length of the arc but does not depend on its position. Thus, equally probable events are stops of the rod in any arcs of the circle that have the same length. After such a simple calculation, it becomes quite obvious that the definitions of probability in the first and second solutions are discrepant. According to the first solution, the probability that the rod will stop in the interval from A to x is $\frac{x}{D}$. The probability that the projection of the point of intersection of the rod with the circle in the second solution will fall in the same interval is, as elementary geometric calculations show, equal to

$$\frac{1}{\pi} \arccos \frac{D-2x}{D} \text{ for } x \leq \frac{D}{2}$$

and

$$1 - \frac{1}{\pi} \arccos \frac{2x-D}{D} \text{ for } x \geq \frac{D}{2}$$

Finally, in the third solution we throw a point into the circle at random and ask ourselves about the probability of its falling within a certain smaller concentric circle (Fig. 4c).

The different statements of the problems in all three cases are quite obvious.

Example 3. Buffon's Needle Problem. A plane is partitioned by parallel lines separated by a distance of $2a$. A needle of length $2l$ ($l < a$) is thrown at random* onto the plane. Find the probability of the needle lying athwart one of the lines.

Solution. Denote by x the distance from the centre to the closest parallel and by φ the angle formed by the needle and this parallel. The quantities x and φ fully define the position of the needle. All possible positions of the needle are defined by points of a rectangle with sides a and π . From Fig. 5 it is seen that for the needle to cross one of the parallel lines it is necessary and sufficient that

$$x \leq l \sin \varphi$$

The sought-for probability is, by the assumptions that have been made, equal to the ratio of the area shaded in Fig. 6 to the area of the rectangle:

$$p = \frac{1}{a\pi} \int_0^\pi l \sin \varphi d\varphi = \frac{2l}{a\pi}$$

* Here, "at random" implies that, firstly, the centre of the needle falls at random on a segment of length $2a$ perpendicular to the drawn straight lines, secondly, the probability that the angle φ made by the needle and the drawn lines will lie between φ_1 and $\varphi_1 + \Delta\varphi$ is proportional to $\Delta\varphi$ and, thirdly, that the quantities x and φ are independent (see Sec. 9).

It will be noted that Buffon's needle problem served as the starting point for solving certain problems in the theory of gunfire that take shell sizes into account.

The formula obtained was employed for an experimental determination of the approximate value of the number π . A large number of needle-throwing experiments have been carried out. A few are listed below.

Experimenter	Year	Number of throws	Experimental value
Wolf	1850	5000	3.1596
Smith	1855	3204	3.1553
Fox	1894	1120	3.1419
Lazzarini	1901	3408	3.1415929

Since from the formula we obtained there follows the equation

$$\pi = \frac{2l}{ap}$$

for a large number n of throws π is approximately equal to

$$\pi \approx \frac{2ln}{am}$$

where m is the number of intersections obtained.

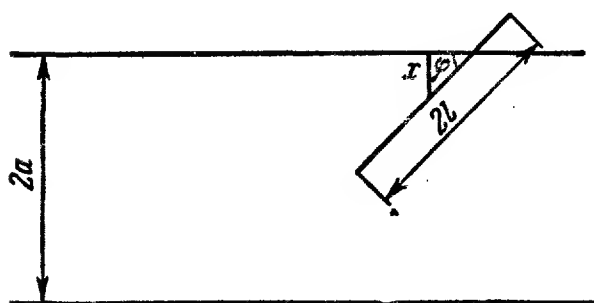


Fig. 5

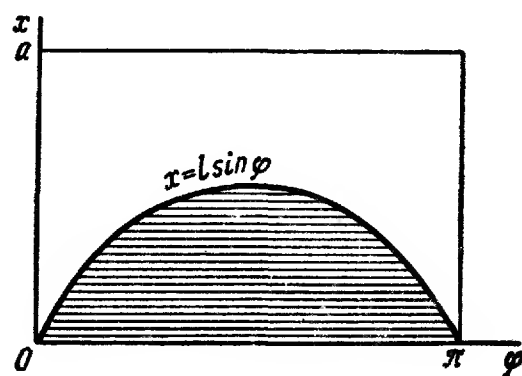


Fig. 6

It will be noted that the results of Fox and Lazzarini are unreliable. In the experiment of the latter, the value of π is given to six exact decimal places. Changing the number of intersections (the number m) by unity affects at least the fourth decimal if n is less than 5000. Indeed, since $a \geq l$,

$$\frac{a(m+1)}{2ln} - \frac{am}{2ln} = \frac{a}{2ln} \geq \frac{1}{2n} \geq 0.0001$$

There is consequently only one value of m which could give the value of π found by Lazzarini. As we shall see in Chapter 2, the probability of obtaining exactly m intersections may be calculated approximately from the formula

$$P_n(m) \approx \frac{1}{\sqrt{2\pi np(1-p)}} e^{-\frac{(m-np)^2}{2np(1-p)}}$$

If for the sake of definiteness we assume that $a=2l$, then in the Lazzarini experiments we find for any m that

$$P_n(m) \leq \frac{1}{\sqrt{2\pi np(1-p)}} \approx 0.03$$

Thus, the probability of obtaining Lazzarini's result is less than $1/30$.

Example 4. Throw at random* a convex contour of diameter less than $2a$ on a horizontal plane partitioned by parallel lines separated by a distance of $2a$. Find the probability that the contour will intersect one of the parallel lines.

Solution. First suppose that the convex contour is an n -sided polygon. Let the sides be numbered from 1 to n . If the polygon intersects a parallel line, then this intersection should occur with two of the sides. Denote by $p_{ij}=p_{ji}$ the probability that the intersection will occur with the i th and j th sides. Obviously, an event A which consists in the thrown polygon intersecting one of the parallel lines may be represented in the form of the following sum of pairwise mutually exclusive events:

$$A = (A_{12} + A_{13} + \dots + A_{1n}) + (A_{23} + A_{24} + \dots + A_{2n}) + \dots \\ \dots + (A_{n-2, n-1} + A_{n-2, n}) + A_{n-1, n}$$

where A_{ij} ($i < j$, $i=1, 2, \dots$; $j=1, 2, \dots$) is an event which consists in an intersection of the i th and j th sides with the parallel line. By the addition theorem for probabilities

$$p = P(A) = [P(A_{12}) + P(A_{13}) + \dots + P(A_{1n})] + \\ + [P(A_{23}) + \dots + P(A_{2n})] + \dots + P(A_{n-1, n}) = \\ = (p_{12} + p_{13} + \dots + p_{1n}) + (p_{23} + p_{24} + \dots + p_{2n}) + \dots + p_{n-1, n}$$

* Here, "at random" means that we take any segment rigidly fixed to the curve and throw it at random in the meaning of the preceding example.

It is easy to demonstrate that the notion "at random" thus defined is independent of the choice of the given segment.

Taking advantage of the equation $p_{ij} = p_{ji}$, we can write the probability p in a different way:

$$p = \frac{1}{2} [(p_{12} + p_{13} + \dots + p_{1n}) + (p_{21} + p_{23} + \dots + p_{2n}) + \dots \\ \dots + (p_{n1} + p_{n2} + \dots + p_{n, n-1})]$$

But the sum $\sum_{j=1}^n p_{ij}$ where we put $p_{ii} = 0$ is the probability of the intersection of the i th side of the polygon with one of the parallel lines. If the length of the i th side is denoted by $2l_i$, then from Buffon's problem we find that

$$\sum_{j=1}^n p_{ij} = \frac{2l_i}{\pi a}$$

and, consequently,

$$p = \frac{\sum_{i=1}^n 2l_i}{2\pi a}$$

Denoting by $2s$ the perimeter of the polygon, we obtain

$$p = \frac{s}{\pi a}$$

We thus see that the probability p does not depend either on the number of sides or on the lengths of the sides of the polygon. From this we conclude that the formula that was found holds for any convex contour, for it can always be considered as the limit of a convex polygon with the number of sides increasing to infinity.

Sec. 7. Frequency and Probability

When passing from elementary cases to complex problems, especially those dealt with in natural science or technology, the classical definition of probability encounters insuperable difficulties of a fundamental nature. First of all, in most cases the question arises of the possibility of finding a reasonable way of isolating the "equally probable cases". For example, for reasons of symmetry (on which our arguments are based concerning the equiprobability of events), it appears at present to be at least difficult to derive the probability of decay of an atom of a radioactive substance within a given time interval or to determine the probability that a child which is to be born will be a boy.

Prolonged observations of the occurrence or nonoccurrence of an event A for a large number of repeated trials that occur under an invariable set of conditions \mathfrak{S} show that for a broad range of phenomena the number of occurrences or nonoccurrences of A obeys stable laws.

Namely, if we denote by μ the number of occurrences of an event A in n independent trials, it will be found that the ratio $\frac{\mu}{n}$ for sufficiently large n in the majority of such series of observations will be almost constant, large deviations being progressively rarer as the number of trials is increased.

This kind of *stability of frequencies* (i.e., of the ratios $\frac{\mu}{n}$) was first noted in phenomena of a demographic nature. In antiquity it was already noticed that for whole states and for large cities the ratio of the number of boys born to the total number of births remained unchanged from year to year. In ancient China, in 2238 B. C. this number was, on the basis of censuses, taken to be equal to $\frac{1}{2}$. Later, particularly in the 17th and 18th centuries, a number of fundamental studies were devoted to the statistics of population. It was found that apart from the stability in births of boys and girls there were observed stable regularities of a different character: the percentage of deaths in a definite age bracket for specific groups of the population (of a particular economic and social background), the distribution of persons (of one sex, age and nationality) as to height, breadth of chest, length of footstep, etc.

Laplace, in his book *Essai philosophique sur les probabilités*, relates of a very indicative episode that occurred when he was studying the regularities of birth of boys and girls. Extensive statistical materials that he had studied dealing with London, St. Petersburg, Berlin, and all of France yielded almost exactly coincident ratios of the number of births of boys to the total number of births. Over many decades all these ratios fluctuated about one and the same number, approximately equal to $\frac{22}{43}$. Yet a study of similar statistical materials of Paris for the 40 years between 1745 and 1784 produced a slightly different number $\frac{25}{49}$. Laplace was intrigued by such a substantial difference and he began to search for a rational explanation. A detailed study of the archives showed that the total number of births in Paris included all foundlings. It also came to light that the surrounding population had a preference for abandoning infants of one sex. This social phenomenon was at that time so common that it had appreciably distorted the true picture of births in Paris. When Laplace eliminated the foundlings from the total number of births, it was found that for Paris as well the ratio of boy births to the total number of births was likewise stable and that it was close to the number $\frac{22}{43}$, which is the same for other peoples and for France as a whole.

Since Laplace's time, extensive statistical material has accumulated that permits very confident predictions to be made of the quantitative characteristics of socially important demographic phenomena. In conclusion, we give rather recent statistical findings of a nearly constant frequency for a large number of trials: the distribution of

newborn infants as to sex by month (see Table 3). The findings are taken from H. Cramér's book *Mathematical Methods of Statistics* and represent the official data of Swedish statistics for 1935.

Figure 7 shows the deviations of the frequency of girl births by month from the frequency of girl births for the year. We see that the frequency fluctuates about the number 0.482.

It turns out that for those cases to which the classical definition of probability is applicable, the fluctuation of frequency occurs about the probability of the event p .

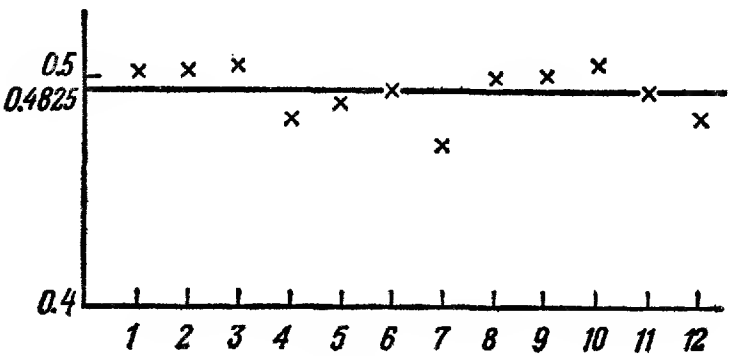


Fig. 7

There is extensive experimental material to verify this fact. Coin tossing, die throwing and needle dropping have all been used to determine empirically the number π (see Example 3 of Sec. 6), and other things. We give some of the results obtained in coin tossing.

Experimenter	Number of throws	Number of "heads"	Frequency
Buffon	4040	2048	0.5080
Karl Pearson	12,000	6019	0.5016
Karl Pearson	24,000	12,012	0.5005

At the present time there are other verifications of this empirical fact which are of important scientific and applied value. In modern statistics, an important role is played by tables of random numbers in which each number is chosen at random from the set of digits 0, 1, 2, . . . , 9. In one of the tables the number 7 appears 968 times in the first 10,000 random numbers, which gives it a frequency of 0.0968 (the probability of occurrence of 7 is equal to 0.1). Computing the number of occurrences of 7 in a sequence of one thousand random numbers, we get the following:

No. of thousand . . .	1	2	3	4	5	6	7	8	9	10
Number of sevens . .	95	88	95	112	95	99	82	89	111	102
Frequency0.095	.088	.095	.112	.095	.099	.082	.089	.111	.102

TABLE 3

Month	1	2	3	4	5	6	7	8	9	10	11	12	One year
Total births	7280	6957	7883	7884	7892	7609	7585	7393	7203	6903	6552	7132	88,273
Boys	3743	3550	4017	4173	4117	3944	3964	3797	3712	3512	3392	3761	45,682
Girls	3537	3407	3866	3711	3775	3665	3621	3596	3491	3391	3160	3371	42,591
Frequency of girls	0.486	0.489	0.490	0.471	0.478	0.482	0.462	0.484	0.485	0.491	0.482	0.473	0.4825

The frequencies of occurrence of a seven in the different thousands fluctuate rather considerably, but are still comparatively close to the probability.

The fact that in a large number of trials the frequency of a series of random events remains almost constant compels us to presume that there are regularities governing the course of the phenomenon which are independent of the investigator, the manifestation of which regularities is this nearly constant frequency. Again, the fact that the frequency of events, to which the classical definition of probability is applicable, is as a rule (in a large number of trials) close to the probability compels us to the view that in the general case there is some constant about which the frequency fluctuates. It is natural to term this constant, which is an objective numerical characteristic of the phenomenon, the *probability* of the random event *A* under study.

We will thus say that event *A* has a probability if it possesses the following peculiarities:

- (a) it is possible, at least theoretically, to carry out under the same conditions ∞ an unlimited number of independent trials, in each of which *A* may or may not occur;
- (b) as a result of a sufficiently large number of trials it is noted that the frequency of the event *A* for nearly every large run of trials departs only slightly from a certain (generally speaking, unknown) constant.

In a large number of trials, the numerical value of this constant may approximately be taken to be the frequency of the event *A* or a number close to the frequency. The probability of a random event thus defined is called *statistical probability*.

Note that frequency has the following properties:

- (1) the frequency of an event that is certain is unity;

- (2) the frequency of the impossible event is zero;
- (3) if a random event C is the sum of mutually exclusive events A_1, A_2, \dots, A_n , then its frequency is equal to the sum of the frequencies with which the component events occur.

Quite naturally, in the case of the statistical definition, we must require that probability satisfy the following properties:

- (1) the probability of an event that is certain is unity;
- (2) the probability of the impossible event is zero;
- (3) if a random event C is the sum of a finite number of mutually exclusive events A_1, A_2, \dots, A_n having probability, then its probability exists and is equal to the sum of the probabilities of the components:

$$P(C) = P(A_1) + P(A_2) + \dots + P(A_n)$$

The statistical definition of probability given here is descriptive rather than formally mathematical in character. It is deficient in yet another aspect as well: it does not lay bare the actual peculiarities of those phenomena for which the frequency is stable. This is to stress the necessity of further investigations in the indicated direction. However, and this is particularly important, in the given definition we retain the objective character of probability that is independent of the investigator. The fact that only after performing certain preliminary observations we can judge that some event has a probability does not in the least detract from our conclusions, for a knowledge of regularities is never derived from nothing; it is always preceded by experiment and observation. Of course, these regularities existed prior to the intervention of the experimenting thinking person, but they were simply unknown to science.

We have already said that we have not given a formally mathematical definition of probability but have only postulated its existence for certain conditions and have indicated a method for an approximate evaluation of it. Any objective property of a phenomenon subjected to study, including the probability of event A , must be determined solely from the structure of the phenomenon, irrespective of whether an experiment is performed or not and whether an experimenting intellect is present or not. Nevertheless, experiment plays an essential role: first of all, it is precisely experiment that enables one to perceive theoretico-probabilistic regularities in nature, secondly, it permits finding in approximate fashion certain probabilities of the events under study, and, finally, it enables us to verify the correctness of the theoretical premises that we use in our investigations. This circumstance requires an explanation.

Suppose that certain arguments suggest that the probability of some event A is p . Further, in a series of independent trials let it be that the frequencies, for the most part, deviate substantially from p . This justifies any doubt we may entertain about the correctness of

our a priori judgements and justifies undertaking a more detailed study of the premises on which our a priori conclusions were based. For instance, we assume that the die we are using has regular geometric forms and that the material it is made of is homogeneous. From these preliminary premises we are entitled to conclude that when throwing the die the probability of any face coming up (say, with number 5) must be equal to $1/6$. If repeated series of large numbers of trials (tosses) systematically demonstrate that the frequency of occurrence of this face departs significantly from $1/6$, then we will not doubt the existence of a definite probability of this face turning up, but will be skeptical about our premises concerning the regularity of the die or proper organization of the trials (tosses).

In conclusion, we must examine the very widespread (particularly among naturalists) conception of probability given by R. von Mises. According to von Mises, since the frequency deviates less and less from the probability p as the number of experiments increases, we should have in the limit

$$p = \lim_{n \rightarrow \infty} \frac{\mu}{n}$$

R. von Mises proposes to regard this equation as a definition of the concept of probability. In his opinion, any a priori definition is doomed and only his own empirical definition is capable of ensuring the interests of natural science, mathematics and philosophy; and since the classical definition has only an extremely limited application, while the statistical definition is applicable to all cases of scientific interest, von Mises proposes rejecting outright the classical definition in terms of equal probability based on symmetry. What is more, Mises considers it quite unnecessary to elucidate the structure of the phenomena for which probability is an objective numerical characteristic, it being regarded as sufficient to have an empirical stability of frequency.

Von Mises believes that the theory of probability has to do with infinite sequences (called *collectives*) of outcomes of trials. Every collective must possess two properties:

(1) the *existence of limiting values* of relative frequencies of those of its members which possess some property of a certain specific group of members;

(2) *randomness*, i.e., invariance of these limits relative to extraction (from the collective) of any subsequence in accordance with a law that is arbitrary except that it must not be based on any difference of the elements of the collective with respect to the property under consideration.

The construction of a mathematical theory based on fulfillment of both these requirements encounters insuperable logical difficulties. The point is that the requirement of randomness is found to be

incompatible with the requirement of the existence of a limit. We shall not dwell on the details of von Mises' theory. We refer the reader to his book *Probability, Statistics, and Truth*. For extensive criticism, see articles by A. Ya. Khinchin*.

It will be noted that observations of statistical stability of frequencies of many actual phenomena served as the starting point for constructing a theory of probability as a mathematical science. The relations that obtain for frequencies serve as the prototype of the principal relations that are satisfied by the probabilities of appropriate events. From this it is clear why the theory of probability may be defined as the domain of mathematics that treats of mathematical models of random phenomena which possess the property of stability of frequencies.

Sec. 8. An Axiomatic Construction of the Theory of Probability

Up until fairly recently, the theory of probability had not established itself as a mathematical science and the basic concepts were not defined with sufficient precision. This vagueness frequently led to paradoxical conclusions (recall the paradoxes of Bertrand). Quite naturally, applications of probability theory to the study of natural phenomena were but feebly substantiated and occasionally encountered sharp and justified criticism. It must be said that these circumstances did not greatly embarrass natural scientists and their naive theoretico-probabilistic approach in various fields of science led to big successes. The development of natural science at the start of this century made stringent demands on probability theory. It became necessary to study systematically the basic concepts of probability theory and to clarify the conditions under which the results of the theory could be employed. That is why a formal-logical substantiation of the theory of probability and its axiomatic construction became so important. In this approach, certain premises, which were a generalization of many centuries of human experience, had to be laid to form the foundation of probability theory as a mathematical science. Its subsequent development must build up via deductions from these basic principles without resort to pictorial conceptions and "common sense" conclusions. In other words, probability theory must be constructed from axioms just like any other established mathematical science such as geometry, theoretical mechanics, abstract group theory, etc.

In modern mathematics, axioms are taken to be propositions that

* "Mises on Probability and the Principles of Physical Statistics", *Uspekhi fizicheskikh nauk*, IX, Issue 2, 1929. "The Frequency Theory of R. von Mises and Modern Ideas of Probability Theory", *Voprosy filosofii*, No. 1, p. 91-102; No. 2, p. 77-89, 1961. Both in Russian.

are regarded as true and are not proved within the framework of the given theory. All other propositions of the theory have to be derived from the accepted axioms in purely logical fashion. Formulation of the axioms (that is, the fundamental propositions on the basis of which an extensive theory is built up) does not represent the initial stage in the development of mathematical science, but is the result of a prolonged accumulation of facts and a logical analysis of the results obtained with the purpose of revealing the actual basic primary facts. That precisely is the way in which the axioms of geometry that are studied in elementary mathematics took shape. The same pathway was taken by probability theory, in which the axiomatic construction of its principles was carried out in comparatively recent times. For the first time, the problem of an axiomatic construction of probability theory as a logically complete science was posed and solved in 1917 by the noted mathematician S. N. Bernstein, who proceeded from a qualitative comparison of random events on the basis of their greater or lesser probability.

A different approach has been proposed by A. N. Kolmogorov, which closely relates the theory of probability with the modern metric theory of functions and also set theory. This book will follow Kolmogorov's approach.

We shall see that the axiomatic construction of the principles of probability theory proceeds from the basic properties of probability noticed in examples of the classical and statistical definitions. Thus, the axiomatic definition of probability includes both the classical and the statistical definitions as particular cases and overcomes the deficiencies of each of them. On this basis it was possible to construct a logically perfect structure of the modern theory of probability and at the same time to satisfy the enhanced requirements of modern natural science.

In Kolmogorov's axiomatics of probability theory the concept of a random event is not primary and is constructed out of more elementary notions. We already encountered that approach when examining certain examples. For instance, in problems dealing with the geometrical definition of probability a region G of space was examined (of a straight line, a plane, etc.) on which a point is thrown at random. Here, the random events are falls in certain subregions of G . Every random event is here a certain subset of the set of points G . This idea underlies the general concept of a random event in the axiomatics of Kolmogorov.

Kolmogorov starts from a set (a space) U of *simple (elementary) events*. The elements of this set are immaterial for the logical development of the theory of probability. The next to be considered is a certain family F of subsets of the set U ; the elements of the family F are called *random events*. It is assumed that the following three requirements are fulfilled relative to the structure of the family F :

- (1) F contains the set U as one of its elements.
- (2) If A and B —subsets of U —are elements of F , then the sets $A+B$, AB , \bar{A} and \bar{B} are also elements of F .

Here, $A+B$ is understood to be a set composed of the elements of U that are components either of A or of B or of A and of B ; by AB is understood the set consisting of the elements of U belonging both to A and B , and, finally, by \bar{A} (\bar{B}), the set of elements of U that do not belong to A (to B).

Insofar as the entire set U belongs to F (as an element), by the second requirement F also contains \bar{U} , that is, F contains the empty set as an element.

It is readily seen that the second requirement implies belonging to the set F of sums, products and complements of a finite number of events belonging to F . Thus, elementary operations on random events cannot take us beyond the limits of the set of random events. As in Sec. 3, we shall call the family of events F a *field of events*.

In many very important problems we shall have to demand more of a field of events, namely:

- (3) If subsets $A_1, A_2, \dots, A_n, \dots$ of set U are elements of the set F , then their sum $A_1+A_2+\dots+A_n+\dots$ and product $A_1A_2\dots A_n\dots$ are also elements of F .

The set F formed in this fashion is called a *Borel field of events* (another now frequently used term is “ σ -algebra of events”).

The foregoing definition of a random event is in full agreement with the picture we got when examining concrete examples. To make this still clearer let us consider two instances in detail from this viewpoint.

Example 1. A die is thrown. The set U of elementary events consists of six elements: $E_1, E_2, E_3, E_4, E_5, E_6$. Here, E_i signifies a roll of i points. The set F of random events consists of the following $2^6=64$ elements: $(V), (E_1), (E_2), (E_3), (E_4), (E_5), (E_6), (E_1, E_2), (E_1, E_3), \dots, (E_5, E_6), (E_1, E_2, E_3), \dots, (E_4, E_5, E_6), (E_1, E_2, E_3, E_4), \dots, (E_3, E_4, E_5, E_6), (E_1, E_2, E_3, E_4, E_5), \dots, (E_2, E_3, E_4, E_5, E_6), (E_1, E_2, E_3, E_4, E_5, E_6)$.

Here, each pair of parentheses indicates which of the elements of set U are used to make up a subset belonging to F (as an element); the symbol (V) indicates the empty set.

Example 2. Encounter Problem. The set U consists of points of the square: $0 \leq x \leq 60, 0 \leq y \leq 60$.

The set F consists of all Borel sets composed of points of this square. In particular, the set consisting of points of the closed region $|x-y| \leq 20$ is contained in F and is a random event.

It is natural to introduce the following definitions.

If two random events A and B do not contain the same elements

of the set U , we shall call them *mutually exclusive*.

The random event U will be called *certain*, and the random event \bar{U} (empty set) the *impossible event*. Events A and \bar{A} are *contrary* (*complementary*) *events*.

Now we can formulate the axioms that define probability.

Axiom 1. *With each random event A in the field of events F there is associated a nonnegative number $P(A)$, called its probability.*

Axiom 2. $P(U)=1$.

Axiom 3 (Axiom of Addition). *If events A_1, A_2, \dots, A_n are pairwise mutually exclusive, then*

$$P(A_1 + A_2 + \dots + A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

For the classical definition of probability there was no need to postulate the properties expressed by Axioms 2 and 3, since these properties of probability were proved by us. And the assertion in Axiom 1 is contained in the classical definition of probability itself.

From these axioms we shall derive several important elementary corollaries. *

First of all, from the obvious equality

$$U = V + \bar{V}$$

and Axiom 3 we conclude that

$$P(U) = P(V) + P(\bar{V})$$

Thus,

1. The probability of the impossible event is zero.

In similar fashion it is easy to detect that

2. For any event A

$$P(\bar{A}) = 1 - P(A)$$

3. No matter what the random event A ,

$$0 \leq P(A) \leq 1$$

4. If event A implies event B , then

$$P(A) \leq P(B)$$

5. Let A and B be two arbitrary events. Insofar as the summands in the sums $A + B = A + (B - AB)$ and $B = AB + (B - AB)$ are mutually exclusive events, in accordance with Axiom 3

$$P(A + B) = P(A) + P(B - AB); \quad P(B) = P(AB) + P(B - AB)$$

* As we shall see in Sec. 29, the teaching of probability is reduced by these axioms to the theory of measure defined on Borel fields of sets. Probability itself is a nonnegative additive set function.

From this follows the addition theorem for arbitrary events A and B :

$$P(A + B) = P(A) + P(B) = P(AB)$$

By virtue of the nonnegativity of $P(AB)$ we conclude that

$$P(A + B) \leq P(A) + P(B)$$

By induction we now derive that if A_1, A_2, \dots, A_n are arbitrary events, we have the inequality

$$P\{A_1 + A_2 + \dots + A_n\} \leq P(A_1) + P(A_2) + \dots + P(A_n)$$

The system of axioms of Kolmogorov is *consistent*, for there exist real objects that satisfy all these axioms. For example, if U is taken as an arbitrary set with a finite number of elements $U = \{a_1, a_2, \dots, a_n\}$, F as a collection of all subsets $\{a_{i_1}, a_{i_2}, \dots, a_{i_s}\}$, $0 \leq i_1 < i_2 < \dots < i_s \leq n$, $0 \leq s \leq n$, then putting

$$P(a_1) = p_1, P(a_2) = p_2, \dots, P(a_n) = p_n$$

where p_1, p_2, \dots, p_n are arbitrary nonnegative numbers that satisfy the equality $p_1 + p_2 + \dots + p_n = 1$, and $P(a_{i_1}, a_{i_2}, \dots, a_{i_s}) = p_{i_1} + \dots + p_{i_s}$, we will satisfy all of Kolmogorov's axioms.

The system of axioms of Kolmogorov is *incomplete*: even for one and the same set U we can choose the probabilities in the set F in different ways.

To take an example, in the case of the die that we examined earlier we can either put

$$P(E_1) = P(E_2) = \dots = P(E_6) = \frac{1}{6} \quad (1)$$

or

$$P(E_1) = P(E_2) = P(E_3) = \frac{1}{4}, P(E_4) = P(E_5) = P(E_6) = \frac{1}{12} \quad (2)$$

and so forth.

The incompleteness of a system of axioms in probability theory is not an indication of an inapt choice or insufficient mental effort in their construction, but is due to the essence of the matter: in various problems there may be phenomena whose study demands consideration of identical sets of random events but with different probabilities. For instance, there may be dice one of which is a true die (exact cube with identical density at every point) and the other not true. In the first case, the system of probabilities will be specified by the system of equations (1), the second, say, by the system (2).

Further development of the theory requires an additional proposition, which is called the *extended axiom of addition*. The new axiom has to be introduced due to the fact that in probability theory one

constantly has to deal with events that decompose into an infinite number of particular cases.

Extended Axiom of Addition. *If an event A is equivalent to the occurrence of at least one of two pairwise mutually exclusive events $A_1, A_2, \dots, A_n, \dots$, then*

$$P(A) = P(A_1) + P(A_2) + \dots + P(A_n) + \dots$$

It will be noted that the extended axiom of addition can be replaced by the *axiom of continuity*, which is equivalent to it.

Axiom of Continuity. *If a sequence of events $B_1, B_2, \dots, B_n, \dots$ is such that each succeeding event implies the preceding event and the product of all events B_n is the impossible event, then*

$$P(B_n) \rightarrow 0 \text{ when } n \rightarrow \infty$$

We shall prove the equivalence of these propositions.

1. The axiom of continuity follows from the extended axiom of addition. Indeed, let events $B_1, B_2, \dots, B_n, \dots$ be such that

$$B_1 \supset B_2 \supset \dots \supset B_n \supset \dots$$

and for any $n \geq 1$

$$\prod_{k \geq n} B_k = V \quad (3)$$

It is obvious that

$$B_n = \sum_{k=n}^{\infty} B_k \bar{B}_{k+1} + \prod_{k \geq n} B_k$$

Since the events in this sum are pairwise mutually exclusive, it follows, according to the extended axiom of addition, that

$$P(B_n) = \sum_{k=n}^{\infty} P(B_k \bar{B}_{k+1}) + P\left(\prod_{k \geq n} B_k\right)$$

But by virtue of condition (3)

$$P\left(\prod_{k \geq n} B_k\right) = 0$$

therefore

$$P(B_n) = \sum_{k=n}^{\infty} P(B_k \bar{B}_{k+1})$$

that is, $P(B_n)$ is the remainder of the convergent series

$$\sum_{k=1}^{\infty} P(B_k \bar{B}_{k+1}) = P(B_1)$$

For this reason $\mathbf{P}(B_n) \rightarrow 0$ as $n \rightarrow \infty$.

2. The extended axiom of addition follows from the axiom of continuity. Let events $A_1, A_2, \dots, A_n, \dots$ be pairwise mutually exclusive and

$$A = A_1 + A_2 + \dots + A_n + \dots$$

We put

$$B_n = \sum_{k=n}^{\infty} A_k$$

It is clear that $B_{n+1} \subset B_n$. If event B_n has occurred, then some one of the events $A_i (i \geq n)$ has occurred and, hence, by virtue of pairwise mutual exclusiveness of events A_k , events A_{i+1}, A_{i+2}, \dots did not occur. Thus, events B_{i+1}, B_{i+2}, \dots are impossible and,

consequently, the event $\prod_{k=n}^{\infty} B_k$ is impossible. By the axiom of continuity, $\mathbf{P}(B_n) \rightarrow 0$ as $n \rightarrow \infty$. Since

$$A = A_1 + A_2 + \dots + A_n + B_{n+1}$$

we have, from the ordinary axiom of addition,

$$\begin{aligned} \mathbf{P}(A) &= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \dots + \mathbf{P}(A_n) + \mathbf{P}(B_{n+1}) = \\ &= \lim_{n \rightarrow \infty} \sum_{k=1}^n \mathbf{P}(A_k) = \sum_{k=1}^{\infty} \mathbf{P}(A_k) \end{aligned}$$

In conclusion we can say that from the point of view of set theory, the axiomatic definition of probability that we have given here is nothing other than the introduction into the set U of a normed, countably additive, nonnegative measure defined on all elements of the set F .

When defining the concept of probability we have to indicate not only the initial set of elementary events U (in modern works it is frequently denoted by the letter Ω as well), but also the set of random events F and the function \mathbf{P} defined on it. The collection $\{U, F, \mathbf{P}\}$ is called a *probability space*.

Sec. 9. Conditional Probability and the Most Elementary Basic Formulas

We have already stated that a certain set of conditions \mathfrak{S} underlies the definition of probability of an event. If no constraints other than the conditions \mathfrak{S} are imposed in computing the probability $\mathbf{P}(A)$, the probability is called *unconditional*.

However, in a number of cases it is necessary to find the probability of events, given the supplementary condition that a certain

event B has occurred that has a positive probability. We shall call such probabilities *conditional* and denote them by the symbol $P(A/B)$; this signifies the probability of event A on condition that event B has occurred. Strictly speaking, unconditional probabilities are also conditional probabilities, since for the starting point of the theory we supposed the existence of a certain invariable set of conditions \mathfrak{S} .

Example 1. Two dice are thrown. What is the probability that the sum 8 comes up (event A) if it is known that this sum is an even number (event B)?

Table 4 gives all possible cases for rolls of two dice. Each cell indicates a possible event: in parentheses, the first number is the sum of the first die, the second, the sum of the second die in each throw.

The total number of possible cases is 36, of which 5 are favourable to event A . Thus, the unconditional probability is

$$P(A) = \frac{5}{36}$$

If event B has occurred, then one of 18 (not 36) possibilities is realized and, consequently, the conditional probability is

$$P(A/B) = \frac{5}{18}$$

Example 2. Two cards are drawn in succession from a deck of cards. Find: (1) the unconditional probability that the second card will be an ace (which card was drawn first is unknown), and (b) the conditional probability that the second card will be an ace if the first was an ace.

TABLE 4

(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

Denote by A the event which consists in the occurrence of an ace in the second place and by B the event consisting in the appearance of an ace in the first place. Clearly, we have the equation

$$A = AB + A\bar{B}$$

By virtue of the fact that the events AB and $A\bar{B}$ are mutually exclusive, we have

$$P(A) = P(AB) + P(A\bar{B})$$

Drawing two cards from a deck of 36 cards can yield $36 \cdot 35$ (taking into account the order!) cases. Of them, there will be $4 \cdot 3$ cases favourable to the event AB and $32 \cdot 4$ cases favourable to the event $A\bar{B}$. Thus,

$$P(A) = \frac{4 \cdot 3}{36 \cdot 35} + \frac{32 \cdot 4}{36 \cdot 35} = \frac{1}{9}$$

If the first card is an ace, there are 35 cards left and only three aces. Hence,

$$P(A/B) = \frac{3}{35}$$

The general solution to the problem of finding conditional probability for the classical definition of probability does not present any difficulty. Indeed, out of n uniquely possible, mutually exclusive and equally probable events A_1, A_2, \dots, A_n let

m events be favourable to event A
 k events be favourable to event B
 r events be favourable to event AB

(naturally, $r \leq k$, $r \leq m$). If event B occurred, this means that one of the events A_j has occurred that is favourable to B . Given this condition, r and only r events A_j favourable to AB are favourable to event A . Thus,

$$P(A/B) = \frac{r}{k} = \frac{\frac{r}{n}}{\frac{k}{n}} = \frac{P(AB)}{P(B)} \quad (1)$$

In exactly the same way we can derive

$$P(B/A) = \frac{P(AB)}{P(A)} \quad (1')$$

Naturally, if $B(A)$ is the impossible event, then Equation (1) (and, respectively, (1')) becomes meaningless.

Each of Equations (1) and (1') is equivalent to the so-called *theorem of multiplication*, according to which

$$P(AB) = P(A)P(B/A) = P(B)P(A/B) \quad (2)$$

that is, *the probability of the product of two events is equal to the product of the probability of one of the events by the conditional probability of the other provided that the first has taken place.*

The multiplication theorem is also applicable in the case when one of the events A and B is the impossible event, since in this case, along with $\mathbf{P}(A)=0$, we have the equations $\mathbf{P}(A/B)=0$ and $\mathbf{P}(AB)=0$.

Conditional probability possesses all the properties of probability. It is easy to see this by checking and finding that it satisfies all the axioms formulated in the preceding section. Indeed, the first axiom is satisfied in obvious fashion, since for each event A a non-negative function $\mathbf{P}(A/B)$ is defined in accordance with (1). If $A=B$, then by the definition (1)

$$\mathbf{P}(B/B) = \frac{\mathbf{P}(BB)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1$$

The third axiom can be verified in the same simple fashion and we leave this to the reader.

It will be noted that the probability space for conditional probabilities is specified by the following triplet $\left\{B, FB, \frac{\mathbf{P}(AB)}{\mathbf{P}(B)}\right\}$.

We say that event A is independent of event B if we have the following equation:

$$\mathbf{P}(A/B) = \mathbf{P}(A) \quad (3)$$

that is, if the occurrence of event B does not alter the probability of event A .

If event A is independent of B , then by virtue of (2) we have the equation

$$\mathbf{P}(A) \mathbf{P}(B/A) = \mathbf{P}(B) \mathbf{P}(A)$$

From this we find:

$$\mathbf{P}(B/A) = \mathbf{P}(B) \quad (4)$$

that is, event B is also independent of A . Thus, *the property of the independence of the events is mutual.*

If events A and B are independent, then events A and \bar{B} are also independent. Indeed, since

$$\mathbf{P}(B/A) + \mathbf{P}(\bar{B}/A) = 1$$

and by assumption $\mathbf{P}(B/A) = \mathbf{P}(B)$, it follows that

$$\mathbf{P}(\bar{B}/A) = 1 - \mathbf{P}(B) = \mathbf{P}(\bar{B})$$

Whence we draw the important conclusion that *if events A and B are independent, then every pair of events (\bar{A}, B) , (A, \bar{B}) , (\bar{A}, \bar{B}) are also independent.*

The notion of the independence of events plays a significant role in probability theory and its applications. Most of the results given in this book have been obtained on the assumption that the events under study are independent.

For a practical determination of the independence of any events, one rarely checks to see that Equations (3) and (4) hold for them. The usual approach is intuition based on experience.

For example, it is clear that a fall of heads in a toss of one coin does not alter the probability of heads (or tails) coming on another coin, if the coins are in no way connected (tied together rigidly, for example) during the tossing. In exactly the same way, the birth of a boy by one mother does not alter the probability of a boy (or a girl) being born by another mother. The events are independent.

For independent events, the *multiplication theorem* takes on a particularly simple form, namely, *if events A and B are independent, then*

$$P(AB) = P(A) \cdot P(B)$$

We shall now generalize the notion of independence of two events to a collection of several events.

Events B_1, B_2, \dots, B_s are called *collectively independent* if for any event B_p of them and arbitrary $B_{i_1}, B_{i_2}, \dots, B_{i_r}$ ($i_n \neq p$) of that same number, events B_p and $B_{i_1}, B_{i_2}, \dots, B_{i_r}$ are mutually independent.

By virtue of the foregoing, this definition is equivalent to the following: for any $1 \leq i_1 < i_2 < \dots < i_r \leq s$ and r ($1 \leq r \leq s$)

$$P(B_{i_1} B_{i_2} \dots B_{i_r}) = P(B_{i_1}) P(B_{i_2}) \dots P(B_{i_r})$$

Note that for several events to be collectively independent it is not sufficient for them to be pairwise independent. This is clear from the following simple example. Suppose the faces of a tetrahedron are coloured as follows: the first red (A), the second green (B), the third blue (C), and the fourth in all three colours (ABC). It is easy to see that the probability of red coming up in a throw of the tetrahedron is equal to $1/2$: there are four faces and two of them have red. Thus,

$$P(A) = \frac{1}{2}$$

In exactly the same way we can calculate that

$$\begin{aligned} P(B) = P(C) = P(A/B) = P(B/C) = P(C/A) = P(B/A) = \\ = P(C/B) = P(A/C) = \frac{1}{2} \end{aligned}$$

Events A, B, C are thus pairwise independent.

However, if we know that events B and C have occurred, then event A has definitely occurred; that is

$$P(A/BC) = 1$$

Thus, events A, B, C are collectively dependent. The above example is due to S. N. Bernstein.

Formula (1'), which in the case of the classical definition was derived by us from the definition of conditional probability, will be taken as a definition in the case of the axiomatic definition of probability. So *in the general case*, for $P(A) > 0$, we have by definition

$$P(B/A) = \frac{P(AB)}{P(A)}$$

(In the case $P(A) = 0$, the conditional probability $P(B/A)$ remains undefined.) This enables us to carry over automatically to the general notion of probability all the definitions and results of the present section.

Now suppose that event B can occur together with one and only one of the n mutually exclusive events A_1, A_2, \dots, A_n . In other words, we put

$$B = \sum_{i=1}^n BA_i \quad (5)$$

where the events BA_i and BA_j with different subscripts i and j are mutually exclusive. By the theorem of the addition of probabilities we have

$$P(B) = \sum_{i=1}^n P(BA_i)$$

Utilizing the multiplication theorem we find

$$P(B) = \sum_{i=1}^n P(A_i) P(B/A_i)$$

This equation is called the *formula of total probability* and plays a basic role throughout the subsequent theory.

By way of illustration we consider two examples.

Example 3. There are five urns:

2 urns of composition A_1 with two white balls and one black ball each,

1 urn of composition A_2 with 10 black balls each,

2 urns of composition A_3 with three white balls and one black ball each.

An urn is selected at random and one ball is drawn from it randomly. What is the probability that the drawn ball will be white (event B)?

Since the ball can be drawn only from urns of the 1st, 2nd, or 3rd composition, it follows that

$$B = A_1B + A_2B + A_3B$$

By the formula of total probability

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)$$

But

$$P(A_1) = \frac{2}{5}, \quad P(A_2) = \frac{1}{5}, \quad P(A_3) = \frac{2}{5},$$

$$P(B/A_1) = \frac{2}{3}, \quad P(B/A_2) = 0, \quad P(B/A_3) = \frac{3}{4}$$

And so

$$P(B) = \frac{2}{5} \cdot \frac{2}{3} + \frac{1}{5} \cdot 0 + \frac{2}{5} \cdot \frac{3}{4} = \frac{17}{30}$$

Example 4. It is known that the probability of receiving k calls at a telephone exchange during a time interval t is equal to $P_t(k)$, ($k=0, 1, 2, \dots$).

Taking it that any number of calls in two adjacent intervals of time are independent events, find the probability of s calls during a time interval of duration $2t$.

Solution. Denote by A_t^k the event consisting of k calls arriving during time t . Obviously, we have the following equation:

$$A_{2t}^s = A_t^0 A_t^s + A_t^1 A_t^{s-1} + \dots + A_t^s A_t^0$$

which means that event A_{2t}^s may be regarded as the sum $s+1$ of mutually exclusive events consisting in the fact that during one interval of time of duration t there are i calls and during the next interval of the same duration there are $s-i$ calls ($i=0, 1, 2, \dots, s$). By the theorem of addition of probabilities,

$$P(A_{2t}^s) = \sum_{i=0}^s P(A_t^i A_t^{s-i})$$

By the theorem of multiplication of probabilities for independent events,

$$P(A_t^i A_t^{s-i}) = P(A_t^i) P(A_t^{s-i}) = P_t(i) \cdot P_t(s-i)$$

Thus, if we put

$$P_{2t}(s) = P(A_{2t}^s)$$

then

$$P_{2t}(s) = \sum_{i=0}^s P_t(i) P_t(s-i) \quad (6)$$

Later on we will see that for certain extremely general conditions ($k=0, 1, 2, \dots$)

$$P_t(k) = \frac{(at)^k}{k!} e^{-at} \quad (7)$$

where a is a certain constant.

From formula (6) we find

$$P_{2t}(s) = \sum_{i=0}^s \frac{(at)^s e^{-2at}}{i! (s-i)!} = (at)^s e^{-2at} \sum_{i=0}^s \frac{1}{i! (s-i)!}$$

But

$$\sum_{i=0}^s \frac{1}{i! (s-i)!} = \frac{1}{s!} \sum_{i=0}^s \frac{s!}{i! (s-i)!} = \frac{1}{s!} (1+1)^s = \frac{2^s}{s!}$$

And so

$$P_{2t}(s) = \frac{(2at)^s e^{-2at}}{s!} \quad (s=0, 1, 2, \dots)$$

Thus, if for a time interval of duration t formula (7) is valid, then for time intervals double that duration and, as it will readily be seen, for any time intervals that are multiples of t , the nature of the formula for the probability continues to hold.

We are now in a position to derive the important *formulas of Bayes* or, as they are sometimes called, "Bayes' rule for the probability of causes", or *Bayes' theorem*, or *Bayes' formulas for the probability of hypotheses*. As before, let Equation (5) be valid. It is required to find the probability of the event A_i if it is known that B has already happened. In accordance with the multiplication theorem we have

$$P(A_i B) = P(B) P(A_i/B) = P(A_i) P(B/A_i)$$

From this

$$P(A_i/B) = \frac{P(A_i) P(B/A_i)}{P(B)}$$

Using the formula of total probability we find

$$P(A_i/B) = \frac{P(A_i) P(B/A_i)}{\sum_{j=1}^n P(A_j) P(B/A_j)}$$

These formulas are due to Bayes. The general procedure for applying them to practical problems is as follows. Let event B take place under diverse conditions, relative to the nature of which one can set up n hypotheses: A_1, A_2, \dots, A_n . For one reason or another we know the probabilities $P(A_i)$ of these hypotheses prior to a trial. It is also known that hypothesis A_i imparts to event B a probability $P(B/A_i)$. An experiment is performed in which B occurs. This should cause a reappraisal of the probabilities of hypotheses A_i ; Bayes' formulas solve this problem quantitatively.

In artillery science we have what is called ranging fire, the purpose of which is to improve our knowledge of the firing conditions (proper aim, for example). Bayes' formula is widely used in the theory of ranging fire. We confine ourselves to a strictly schematic example for the sole purpose of illustrating the type of problems solved by this formula.

Example 5. There are five urns of the following compositions:
 2 urns (composition A_1) with 2 white and 3 black balls each,
 2 urns (composition A_2) with 1 white and 4 black balls each,
 1 urn (composition A_3) with 4 white balls and 1 black ball.

A ball is chosen from one of the urns taken at random. It turned out to be white (event B). What is the probability after the experiment (a posteriori probability) that the ball was taken from the urn of the third composition?

By hypothesis we have

$$P(A_1) = \frac{2}{5}, \quad P(A_2) = \frac{2}{5}, \quad P(A_3) = \frac{1}{5},$$

$$P(B/A_1) = \frac{2}{5}, \quad P(B/A_2) = \frac{1}{5}, \quad P(B/A_3) = \frac{4}{5}$$

By Bayes' formula we have

$$P(A_3/B) = \frac{P(A_3)P(B/A_3)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + P(A_3)P(B/A_3)} =$$

$$= \frac{\frac{1}{5} \cdot \frac{4}{5}}{\frac{2}{5} \cdot \frac{2}{5} + \frac{2}{5} \cdot \frac{1}{5} + \frac{1}{5} \cdot \frac{4}{5}} = \frac{4}{10} = \frac{2}{5}$$

In exactly the same way we find

$$P(A_1/B) = \frac{2}{5}, \quad P(A_2/B) = \frac{1}{5}$$

Sec. 10. Examples

The following are somewhat more complicated examples of the use of the foregoing theory.

Example 1*. Two players A and B continue a certain game to the ultimate ruin of one of them. The capital of the first is a rubles, the capital of the second is b rubles. The probability of winning each play is p for player A and q for player B ; $p+q=1$ (there are no draws). In each play, a win of one player (and, hence, a loss of the other) is equal to one ruble. Find the probability of ruin of each of the players (the outcomes of the separate plays are presumed to be independent).

Solution. Before beginning an analytical solution of the problem let us see what meaning a simple (elementary) event has here and how the probability of the event that interests us is defined.

An elementary event is to be understood as an infinite sequence of alternations of outcomes of the separate plays. For instance, an elementary event (A, \bar{A}, A, \dots) consists in the fact that all odd plays are won by A and all even ones by player B . A random event—the ruin of player A —consists of all elementary events in which A loses his capital before player B does. Note that each elementary event is a countable sequence consisting of letters A and \bar{A} ; for this reason, in each elementary event that enters into the random event (the ruin of player A) that interests us, there will be a countable set of alternations A and \bar{A} after the game culminates in the ruin of player A .

We denote by $p_n(N)$ the probability of ruin of A during N plays if he had n rubles before starting to play. It is easy to determine this probability since the set of elementary events consists solely of a finite number of elements. It is natural here to put the probability of each elementary event equal to $p^m q^{N-m}$, where m is the number of occurrences of A , and $N-m$ is the number of occurrences of \bar{A} in the total number N of occurrences of both letters. In the same way, let $q_n(N)$ and $r_n(N)$ be, respectively, the probabilities of loss of B and a draw after N plays.

It is clear that the numbers $p_n(N)$ and $q_n(N)$ do not diminish with growth of N and the number $r_n(N)$ does not increase. We thus have the limits

$$p_n = \lim_{N \rightarrow \infty} p_n(N), \quad q_n = \lim_{N \rightarrow \infty} q_n(N), \quad r_n = \lim_{N \rightarrow \infty} r_n(N)$$

We shall call these limits the probabilities of loss of players A and B and of a draw, respectively, provided that at the start A

* For this ruin problem we retain the classical formulation; but other formulations are possible, such as: a physical particle lies on a straight line at point O and every second is subjected to a random impulse, as a result of which it is translated 1 cm to the right with probability p or 1 cm to the left with probability $q=1-p$. What is the probability that the particle will find itself to the right of a point with coordinate b ($b > 0$) before it finds itself to the left of a point with coordinate a ($a < 0$, a and b are integers)?

had n rubles and B had $a+b-n$ rubles. Since for any $N > 0$

$$p_n(N) + q_n(N) + r_n(N) = 1$$

it follows that in the limit

$$p_n + q_n + r_n = 1$$

Further, it is obvious that

(1) if at the start player A has all the capital and B has nothing, then

$$p_{a+b} = 0, \quad q_{a+b} = 1, \quad r_{a+b} = 0 \quad (1)$$

(2) if player A has nothing at the beginning, and B has all the capital, then

$$p_0 = 1, \quad q_0 = 0, \quad r_0 = 0 \quad (1')$$

If player A has n rubles prior to some play, then his ruin can occur in two different ways: either he will win the next play and will lose the entire game, or will lose the play and the game. Therefore, from the formula of total probability

$$p_n = p \cdot p_{n+1} + q \cdot p_{n-1}$$

We have a difference equation in p_n ; it is easy to see that we can write it in the following form:

$$q(p_n - p_{n-1}) = p(p_{n+1} - p_n) \quad (2)$$

Let us first examine the solution of this equation for $p = q = \frac{1}{2}$. On this assumption,

$$p_{n+1} - p_n = p_n - p_{n-1} = \dots = p_1 - p_0 = c$$

where c is a constant. From this we find

$$p_n = p_0 + nc$$

Since $p_0 = 1$ and $p_{a+b} = 0$, it follows that

$$p_n = 1 - \frac{n}{a+b}$$

Thus the probability of the ruin of player A is equal to

$$p_a = 1 - \frac{a}{a+b} = \frac{b}{a+b}$$

In like fashion we find that in the case of $p = \frac{1}{2}$ the probability of the ruin of player B is

$$q_a = \frac{a}{a+b}$$

Whence it follows that for $p = q = 1/2$

$$r_a = 0$$

In the general case, for $p \neq q$, we find from (2) that

$$q^n \prod_{k=1}^n (p_k - p_{k-1}) = p^n \prod_{k=1}^n (p_{k+1} - p_k)$$

After simplifications and taking advantage of relations (1), we find

$$p_{n+1} - p_n = \left(\frac{q}{p}\right)^n (p_1 - 1)$$

Consider the difference $p_{a+b} - p_n$; it is obvious that

$$\begin{aligned} p_{a+b} - p_n &= \sum_{k=n}^{a+b-1} (p_{k+1} - p_k) = \sum_{k=n}^{a+b-1} \left(\frac{q}{p}\right)^k (p_1 - 1) = \\ &= (p_1 - 1) \frac{\left(\frac{q}{p}\right)^n - \left(\frac{q}{p}\right)^{a+b}}{1 - \frac{q}{p}} \end{aligned}$$

Since $p_{a+b} = 0$, it follows that

$$p_n = (1 - p_1) \frac{\left(\frac{q}{p}\right)^n - \left(\frac{q}{p}\right)^{a+b}}{1 - \frac{q}{p}}$$

and since $p_0 = 1$, we have

$$1 = (1 - p_1) \frac{\left(\frac{q}{p}\right)^0 - \left(\frac{q}{p}\right)^{a+b}}{1 - \frac{q}{p}}$$

Eliminating p_1 from the last two equations we find

$$p_n = \frac{\left(\frac{q}{p}\right)^{a+b} - \left(\frac{q}{p}\right)^n}{\left(\frac{q}{p}\right)^{a+b} - 1}$$

Hence the probability of ruin of player A is equal to

$$p_a = \frac{q^{a+b} - q^a p^b}{q^{a+b} - p^{a+b}} = \frac{1 - \left(\frac{p}{q}\right)^b}{1 - \left(\frac{p}{q}\right)^{a+b}}$$

In exactly the same way we find that the probability of ruin of player B , for $p \neq q$, is

$$q_b = \frac{1 - \left(\frac{q}{p}\right)^a}{1 - \left(\frac{q}{p}\right)^{a+b}}$$

The last two formulas show that, in the general case, the probability of a draw is zero:

$$r_a = 0$$

From these formulas we can draw the following conclusions: if the capital of one of the players, say B , is incomparably greater than the capital of A , so that for all practical purposes b may be considered infinitely large compared with a , and the players are of equal skill, then the ruin of B is practically impossible. The conclusion will be quite different if A plays better than B and, consequently, $p > q$. Assuming $b \sim \infty$, we find

$$q_a \sim 1 - \left(\frac{q}{p}\right)^a$$

and

$$p_a \sim \left(\frac{q}{p}\right)^a$$

From this we infer that a skilful player with even a slight capital can have less chance of ruin than a player with a large capital but less skilful.

Some problems in physics and technology reduce to the ruin problem.

Example 2. Find the probability that a machine tool operating at time t_0 will not stop till time $t_0 + t$ if it is known that: (1) this probability depends only on the length of the time interval $(t_0, t_0 + t)$; (2) the probability that the machine will stop during time interval Δt is proportional to Δt up to infinitesimals of higher orders* with respect to Δt ; (3) events consisting in machine stoppage in nonoverlapping time intervals are independent.

Solution. We denote the desired probability by $p(t)$. The probability that the machine will stop within the time interval Δt is

$$1 - p(\Delta t) = a\Delta t + o(\Delta t)$$

where a is some constant.

* Henceforward, to state that some quantity α is infinitely small compared to β , we will write $\alpha = o(\beta)$. But if the ratio $\frac{\alpha}{\beta}$ is bounded in absolute value, we shall write $\alpha = O(\beta)$.

Let us determine the probability that the machine, which was in operation at time t_0 , will not stop up to time $t_0 + t + \Delta t$. For this event to happen, it is necessary that the machine should not stop during times of duration t and Δt ; by virtue of the multiplication theorem we thus have

$$p(t + \Delta t) = p(t) \cdot p(\Delta t) = p(t)(1 - a\Delta t - o(\Delta t))$$

And from this we have

$$\frac{p(t + \Delta t) - p(t)}{\Delta t} = -ap(t) - o(1) \quad (3)$$

Let us now pass to the limit, putting $\Delta t \rightarrow 0$; from the fact that the right-hand side of (3) has a limit it follows that the left-hand side has a limit too. And so we find

$$\frac{dp(t)}{dt} = -ap(t)$$

The solution of this equation is the function

$$p(t) = Ce^{-at}$$

where C is a constant. This constant is found from the obvious condition that $p(0) = 1$. Thus*

$$p(t) = e^{-at}$$

The first condition of the problem imposes great restrictions on the operating regime of the machine; however, there are places where it is fulfilled to a high degree of accuracy. An instance is the operation of an automatic loom. We note that many other problems reduce to the one we have considered, such, for example, as that of the distribution of probabilities of the mean free path of a molecule in the kinetic theory of gases.

Example 3. Mortality tables are often compiled on the following assumptions:

(1) the probability that a certain person will die between time t and $t + \Delta t$ is

$$p(t, t + \Delta t) = a(t)\Delta t + o(\Delta t)$$

where $a(t)$ is a nonnegative continuous function:

(2) it is taken that the death of the given person (or his survival) during the time interval (t_1, t_2) under consideration does not depend on what preceded t_1 ;

* Changing the reasoning, we can prove that the result obtained will be the same if we do not assume that the second condition of the problem is fulfilled.

(3) the probability of death at the time of birth is zero.

Proceeding from these assumptions, find the probability of death of person A before he reaches the age of t .

Solution. We denote by $\pi(t)$ the probability that A will live to the age t and we compute $\pi(t+\Delta t)$. From the original assumptions we obviously have the equation

$$\pi(t+\Delta t) = \pi(t) \pi(t+\Delta t; t)$$

where $\pi(t+\Delta t; t)$ signifies the probability of living to the age $t+\Delta t$ if A has already reached age t . In accordance with the first and second assumptions

$$\pi(t+\Delta t; t) = 1 - p(t, t+\Delta t) = 1 - a(t) \Delta t - o(\Delta t)$$

Therefore

$$\pi(t+\Delta t) = \pi(t) [1 - a(t) \Delta t - o(\Delta t)]$$

From this we find that $\pi(t)$ satisfies the following differential equation:

$$\frac{d\pi(t)}{dt} = -a(t) \pi(t)$$

Taking into account the third condition of the problem, the solution of this equation will be the function

$$\pi(t) = e^{-\int_0^t a(z) dz}$$

Thus, the probability of dying before the age of t is equal to

$$1 - \pi(t) = 1 - e^{-\int_0^t a(z) dz}$$

Mortality tables for adults are often compiled on the basis of Makeham's formula, according to which

$$a(t) = \alpha + \beta e^{\gamma t}$$

where the constants α , β , γ are positive. * The derivation of this formula is based on the assumption that a grown person can die from causes that have nothing to do with age and from causes depending on age, the probability of death increasing with age in a geometrical progression. Given that supplementary assumption,

$$\pi(t) = e^{-\alpha t - \frac{\beta}{\gamma} (e^{\gamma t} - 1)}$$

Example 4. In modern nuclear physics, the intensity of a particle source is measured with Geiger-Muller counters. A particle entering

* Their value is determined by the conditions under which the group of persons undergoing study live (social conditions, primarily).

the counter generates a discharge in it that lasts time τ , during which the counter does not record any particles entering the counter. Find the probability that the counter will count all particles entering it during time t if the following conditions are fulfilled:

(1) the probability that during time t a total of k particles will enter the counter is independent of the number of particles that entered the counter prior to this time interval;

(2) the probability that during the time interval from t_0 to $t_0 + t$, k particles entered the counter is given by the formula*

$$p_k(t_0, t_0 + t) = \frac{(at)^k e^{-at}}{k!}$$

where a is a positive constant;

(3) τ is a constant quantity.

Solution. Denote by $A(t)$ the event that all particles that entered the counter during time t were counted; by $B_k(t)$ the event that during time t a total of k particles entered the counter.

By virtue of the first condition of the problem, for $t \geq \tau$,

$$\begin{aligned} P\{A(t + \Delta t)\} &= \\ &= P\{A(t)\} P\{B_0(\Delta t)\} + P\{A(t - \tau)\} P\{B_0(\tau)\} P\{B_1(\Delta t)\} + o(\Delta t) \end{aligned}$$

and for $0 \leq t \leq \tau$

$$P\{A(t + \Delta t)\} = P\{A(t)\} P\{B_0(\Delta t)\} + P\{B_0(t)\} P\{B_1(\Delta t)\} + o(\Delta t)$$

For the sake of brevity, put $\pi(t) = P\{A(t)\}$; then on the basis of the second and third conditions of the problem, for $0 \leq t \leq \tau$,

$$\pi(t + \Delta t) = \pi(t) e^{-a\Delta t} + e^{-a\Delta t} a\Delta t e^{-at} + o(\Delta t)$$

and for $t \geq \tau$

$$\pi(t + \Delta t) = \pi(t) e^{-a\Delta t} + \pi(t - \tau) e^{-a\Delta t} a\Delta t e^{-a\tau} + o(\Delta t)$$

By passing to the limit as $\Delta t \rightarrow 0$ we find that for $0 \leq t \leq \tau$ we have the equation

$$\frac{d\pi(t)}{dt} = -a\pi(t) + ae^{-at} \quad (4)$$

and for $t \geq \tau$, the equation

$$\frac{d\pi(t)}{dt} = -a[\pi(t) - \pi(t - \tau)e^{-a\tau}] \quad (5)$$

* It will become clear later on why in this example and in Example 4 of the previous section we considered that

$$p_k = \frac{(at)^k e^{-at}}{k!}$$

From equation (4) we find that when $0 \leq t \leq \tau$

$$\pi(t) = e^{-at}(c + at)$$

From the condition

$$\pi(0) = 1$$

we determine the constant c . Finally, when $0 \leq t \leq \tau$,

$$\pi(t) = e^{-at}(1 + at) \quad (6)$$

For $\tau \leq t \leq 2\tau$, the probability $\pi(t)$ is determined from the equation

$$\begin{aligned} \frac{d\pi(t)}{dt} &= -a[\pi(t) - \pi(t - \tau)e^{-a\tau}] = \\ &= -a[\pi(t) - e^{-a(t-\tau)}(1 + a(t - \tau))e^{-a\tau}] = \\ &= -a[\pi(t) - e^{-at}(1 + a(t - \tau))] \end{aligned}$$

The solution of this equation gives us

$$\pi(t) = e^{-at} \left(c_1 + at + \frac{a^2(t - \tau)^2}{2!} \right)$$

The constant c_1 may be found from the fact that according to (6)

$$\pi(\tau) = e^{-a\tau}(1 + a\tau)$$

Thus, $c_1 = 1$ and for $\tau \leq t \leq 2\tau$

$$\pi(t) = e^{-at} \left[1 + at + \frac{a^2(t - \tau)^2}{2!} \right]$$

By the method of complete induction it may be proved that for $(n-1)\tau \leq t \leq n\tau$, the following equation is valid:

$$\pi(t) = e^{-at} \sum_{k=0}^n \frac{a^k [t - (k-1)\tau]^k}{k!}$$

EXERCISES

A, B, C are random events.

1. What meaning do the following equations have:

- (a) $ABC = A$;
- (b) $A + B + C = A$?

2. Simplify the expressions

- (a) $(A + B)(B + C)$;
- (b) $(A + B)(A + \bar{B})$;
- (c) $(A + B)(A + \bar{B})(\bar{A} + B)$.

3. Prove the equations:

- (a) $\overline{\overline{AB}} = A + B$;
- (b) $\overline{\overline{A} + \overline{B}} = AB$;
- (c) $\overline{A_1 + A_2 + \dots + A_n} = \overline{A_1} \overline{A_2} \dots \overline{A_n}$;
- (d) $\overline{A_1 A_2 \dots A_n} = \overline{A_1} + \overline{A_2} + \dots + \overline{A_n}$.

4. A four-volume work is on a shelf in random order. What is the probability that the volumes stand in the proper order from left to right or from right to left?

5. The numbers 1, 2, 3, 4, 5 are written on five cards. Three cards are drawn in succession and at random from the deck; the resulting digits are written from left to right. What is the probability that the resulting three-digit number will be even?

6. There are M defective items in a lot consisting of N items. From this lot we select n ($n < N$) items at random. What is the probability that there will be m defective items ($m \leq M$) among them?

7. A quality control inspector examines items in a lot consisting of m items of first grade and n second-grade items. An inspection of the first b items chosen at random from the lot showed that they are all of grade two ($b < m$). What is the probability that of the next two randomly chosen unchecked items at least one will also be of grade two?

8. Using probabilistic arguments, prove the identity ($A > a$):

$$1 + \frac{A-a}{A-1} + \frac{(A-a)(A-a-1)}{(A-1)(A-2)} + \dots + \frac{(A-a)\dots 2 \cdot 1}{(A-1)\dots(a+1)a} = \frac{A}{a}$$

Hint. An urn has A balls, of which a are white. The balls are drawn at random without replacement. Find the probability that sooner or later a white ball will be encountered.

9. Draw one ball after another in succession from a box containing m white balls and n black balls ($m > n$). What is the probability that there will come a time when the number of selected black balls will be equal to the number of white ones drawn?

10. A person wrote letters to n addressees, one letter in each envelope, and then, at random, wrote one of the n addresses on each envelope. What is the probability that at least one of the letters reached its destination?

11. An urn has n tickets with numbers from 1 to n . The tickets are drawn at random, one at a time (without replacement). What is the probability that at least in one selection the number of the extracted ticket will coincide with the number of the trial?

12. From an urn containing n white balls and n black ones select at random an even number of balls (all the different ways of drawing an even number of balls are considered equally probable, irrespective of their number). Find the probability that there will be the same number of black and white balls among them.

13. *The paradox of de Méré.* What is more probable: to get one ace with four dice, or to get one double ace in 24 throws of two dice?

14. Three points are thrown at random on a segment $(0, a)$. Find the probability that a triangle can be constructed out of line-segments equal to distances from point 0 to the points of fall.

15. A rod of length l is broken at two randomly chosen points. What is the probability that the pieces can be used to build a triangle?

16. A point is dropped at random onto line-segment AB of length a . Another point is dropped at random on a line-segment BC of length b . What is the probability that a triangle can be built from the lines: (1) from point A to the

first point; (2) between the two points that were dropped; (3) from the second dropped point to point C ?

17. A total of N points are dropped at random and independently of one another into a sphere of radius R .

(a) What is the probability that the distance from the centre to the nearest point will be at least r ?

(b) What does the probability found in (a) approach if $R \rightarrow \infty$ and $\frac{N}{R^3} \rightarrow \frac{4}{3} \pi \lambda$?

Note. The problem is taken from stellar astronomy: in the vicinity of the sun, $\lambda \approx 0.0063$ if R is measured in parsecs.

18. The events A_1, A_2, \dots, A_n are independent; $P(A_k) = p_k$. Find the probability of:

(a) the occurrence of at least one of these events;

(b) the nonoccurrence of all these events;

(c) the occurrence of exactly one event (it is immaterial which).

19. Prove that if events A and B are mutually exclusive, $P(A) > 0$ and $P(B) > 0$, then events A and B are dependent.

20. Let A_1, A_2, \dots, A_n be random events. Prove the formula

$$P\left\{\sum_{k=1}^n A_k\right\} = \sum_{i=1}^n P(A_i) - \sum_{1 \leq i < j \leq n} P(A_i A_j) + \sum_{1 \leq i < j < k \leq n} P(A_i A_j A_k) \mp \dots \pm P(A_1 A_2 \dots A_n)$$

Employing this formula, solve Problems 10 and 11.

21. The probability that a molecule that collided with another molecule at time $t=0$ and that did not experience any other collisions up to time t will experience a collision during the time interval between t and $t+\Delta t$ is equal to $\lambda \Delta t + o(\Delta t)$. Find the probability that the time of free motion (the time between two successive collisions) will be greater than t .

22. Assuming that in the multiplication of bacteria by fission (division into two bacteria) the probability of a bacterium dividing during a time interval Δt is equal to $a \Delta t + o(\Delta t)$ and is not dependent on the number of previous divisions or on the number of bacteria present, find the probability that if at time 0 there was one bacterium there will be i bacteria at time t .

CHAPTER 2

Sequences of Independent Trials

Sec. 11. Independent Trials. Bernoulli's Formulas

In this chapter we take up the study of basic regularities involved in one of the most important schemes of probability theory, that of a sequence of independent trials. To this notion we ascribe the following meaning.

By a trial we mean the realization of a specific set of conditions that can give rise to an elementary event of a space U of elementary events (sample space). The mathematical model of a sequence of n trials is a new sample space U_n consisting of points (e_1, e_2, \dots, e_n) where e_i is an arbitrary point of the space U corresponding to a trial with the number i .

Suppose the trial consists in tossing a die. The space of elementary states consists of 6 points. The space U_3 which corresponds to three trials consists of 216 points (e_1, e_2, e_3) .

Suppose a trial is regarded as a check of the duration of faultless operation of a semiconductor device at a given voltage. The sample space consists of the set of points of the half-line $0 \leq e < \infty$. The space U_n consists of the set of points (e_1, e_2, \dots, e_n) whose coordinates assume nonnegative values equal to the durations of faultless operation of the devices numbered 1, 2, \dots , n , respectively.

Let us assume that for the s th trial, the space U is partitioned into k mutually exclusive random events $A_1^{(s)}, A_2^{(s)}, \dots, A_k^{(s)}$, that is, we assume that

$$A_1^{(s)} + A_2^{(s)} + \dots + A_k^{(s)} = U, \quad A_i^{(s)} A_j^{(s)} = \emptyset$$

($i \neq j$; $i, j = 1, 2, \dots, k$; $s = 1, 2, \dots, n$). Event $A_i^{(s)}$ will be called the i th outcome in the s th trial. We denote the probability of the i th outcome in the s th trial by $p_i^{(s)} = \mathbf{P}(A_i^{(s)})$.

We denote by $A_{i_1}^{(1)}, A_{i_2}^{(2)}, \dots, A_{i_n}^{(n)}$ an event consisting of all the points (e_1, e_2, \dots, e_n) of space U_n for which $e_1 \in A_{i_1}^{(1)}, e_2 \in A_{i_2}^{(2)}, \dots, e_n \in A_{i_n}^{(n)}$.

If in the space U_n the equation

$$\mathbf{P} \{A_{i_1}^{(1)} A_{i_2}^{(2)} \dots A_{i_n}^{(n)}\} = p_{i_1}^{(1)} p_{i_2}^{(2)} \dots p_{i_n}^{(n)}$$

is valid for any i_1, i_2, \dots, i_n ($1 \leq i_1 < i_2 < \dots < i_n \leq k$), the trials are termed *independent*.*

In the future we shall confine ourselves to the case where the probabilities of events $A_i^{(s)}$ do not depend on the number of the trial; then we denote $p_i = \mathbf{P} \{A_i^{(s)}\}$ ($i=1, 2, \dots, k$); since the outcomes $A_i^{(s)}$ are mutually exclusive and exhaustive, it is obvious that we have $\sum p_i = 1$. This scheme was first considered by James Bernoulli in the highly important special case of $k=2$. The case of $k=2$ is therefore known as the *Bernoulli scheme*. Ordinarily, in the Bernoulli scheme, $p_1 = p$, $p_2 = 1 - p = q$.

The definition of independent trials gives the following result:

Theorem. *If n given trials are independent, then any m of them are also independent.*

For the sake of simplicity, we confine ourselves to the case of $m = n - 1$, since there are no difficulties in passing to the general case. Indeed, we have the obvious equality

$$A_{i_1}^{(1)} A_{i_2}^{(2)} \dots A_{i_{n-1}}^{(n-1)} \sum_{j=1}^k A_j^{(n)} = A_{i_1}^{(1)} A_{i_2}^{(2)} \dots A_{i_{n-1}}^{(n-1)}$$

from which it follows that

$$\mathbf{P} \{A_{i_1}^{(1)} A_{i_2}^{(2)} \dots A_{i_{n-1}}^{(n-1)}\} = \prod_{s=1}^{n-1} \mathbf{P} \{A_{i_s}^{(s)}\} \sum_{j=1}^k \mathbf{P} \{A_j^{(n)}\} = \prod_{s=1}^{n-1} \mathbf{P} \{A_{i_s}^{(s)}\}$$

By definition this means that the first $n - 1$ trials are independent.

It is easy to prove the following theorem, which elucidates the conditions of independence of trials.

Theorem. *For n trials to be independent it is necessary and sufficient to satisfy the conditions*

$$\mathbf{P} \{A_q^{(i)} / A_{q_1}^{(i_1)} \dots A_{q_m}^{(i_m)}\} = \mathbf{P} \{A_q^{(i)}\}$$

for any group of numbers i, i_1, \dots, i_m ($1 \leq i, i_1, \dots, i_m \leq n$) and any m, q, q_1, \dots, q_m ($1 \leq m \leq n, 1 \leq q, q_1, \dots, q_m \leq k$).

We shall not take up the proof of this proposition here, partly because its practical verification involves great difficulties.

* It is possible to ascribe a broader meaning to the scheme of a sequence of independent trials and consider that the number of outcomes and the probabilities of these outcomes depend on the trial number. These more general constructions will not be considered.

A detailed investigation of such sequences of trials deserves the fullest attention both because of their immediate value in probability theory and in applications and also by virtue of the possibility, revealed in the process of the development of probability theory, of generalizing the regularities first discovered in studies of the scheme of a sequence of independent trials, in particular the Bernoulli scheme. Many of the facts elicited in this special scheme were later to serve as a guide in the study of more sophisticated schemes. This remark refers both to the past and to the modern development of probability theory, which will become evident later on in discussing examples of the law of large numbers and the DeMoivre-Laplace theorem.

The most elementary problem involving a scheme of independent trials consists in determining the probability $P_n(m)$ that in n trials an event A will occur m times, and that in the remaining $n - m$ trials the contrary event A will occur.

We first find the probability that events $A^{(s)}$ occur in m specific trials (for instance, in trials with the numbers s_1, s_2, \dots, s_m) and do not occur in the remaining $n - m$ trials. By the multiplication theorem for independent events, this probability is

$$p^m q^{n-m}$$

By the theorem of addition of probabilities, the desired probability $P_n(m)$ is equal to the sum of the above computed probabilities for all the different modes m of occurrence of the event and $n - m$ nonoccurrences from among n trials. Combinatorial theory states that the number of such ways is $C_n^m = \frac{n!}{m!(n-m)!}$; consequently, the probability sought for is

$$P_n(m) = C_n^m p^m q^{n-m} \quad (1)$$

Since all possible mutually exclusive outcomes of n trials consist in the occurrence of event $A^{(s)}$ zero times, 1 time, 2 times, \dots , n times, it is clear that

$$\sum_{m=0}^n P_n(m) = 1$$

This relationship can also be derived, without probabilistic reasoning, from the equation

$$\sum_{m=0}^n P_n(m) = (p + q)^n = 1^n = 1$$

It will readily be seen that the probability $P_n(m)$ is equal to the coefficient of x^m in the binomial expansion of $(q + px)^n$ in powers of

x ; by virtue of this property the collection of probabilities $P_n(m)$ is called the *law of binomial probability distribution*.

By modifying our reasoning somewhat, the reader will easily see that if, in each of the trials, one of k mutually exclusive events A_i can occur and the probability of occurrence of event A_i in each trial is p_i , the probability of occurrence of event A_1 , m_1 times in the course of n trials, of event A_2 , m_2 times, and of event A_k , $(m_1 + m_2 + \dots + m_k = n)$, \dots , m_k times is

$$P_n(m_1, m_2, \dots, m_k) = \frac{n!}{m_1! m_2! \dots m_k!} p_1^{m_1} p_2^{m_2} \dots p_k^{m_k} \quad (1')$$

It is also easy to see that this probability is the coefficient of $x_1^{m_1} x_2^{m_2} \dots x_k^{m_k}$ in the expansion of the *polynomial* $(p_1 x_1 + p_2 x_2 + \dots + p_k x_k)^n$ in powers of x .

Having in view the formulation of general problems involving the independent-trial scheme, we shall consider some numerical examples. We will not work out the desired probabilities to the end but will leave them until convenient methods have been prepared.

Example 1. There are two vessels A and B , each with a volume of 1 dm^3 . Each contains 2.7×10^{22} molecules of gas. The vessels are brought into contact so that there is a free exchange of molecules between them. What is the probability that after 24 hours one of the vessels will have at least one ten-thousand millionth part more molecules than the other?

For each molecule, the probability of being in one or the other vessel 24 hours later is the same and is one half. Thus, it is as if 5.4×10^{22} trials were performed, for each of which the probability of being in vessel A is equal to $1/2$. Let μ be the number of molecules that go to vessel A and, hence, $5.4 \times 10^{22} - \mu$ is the number of molecules that go to vessel B . We have to determine the probability that

$$|\mu - (5.4 \times 10^{22} - \mu)| \geq \frac{5.4 \times 10^{22}}{10^{10}} = 5.4 \times 10^{12}$$

in other words, we must find the probability

$$P = P \{ |\mu - 2.7 \times 10^{22}| \geq 2.7 \times 10^{12} \}$$

By the addition theorem

$$P = \sum P \{ \mu = m \}$$

where the sum is extended to those values of m for which

$$|m - 2.7 \times 10^{22}| \geq 2.7 \times 10^{12}$$

Example 2. The probability that an item of a certain kind of production will be defective is equal to 0.005. What is the proba-

bility that out of 10,000 randomly chosen items there will be (a) exactly 40, (b) no more than 70 defective ones?

In our example, $n = 10,000$, $p = 0.005$. Therefore, by formula (1) we find

$$(a) \quad P_{10,000}^{(40)} = C_{10,000}^{40} (0.995)^{9,960} (0.005)^{40}$$

The probability $P\{\mu \leq 70\}$ that the number of defective items will not turn out to be more than seventy is equal to the sum of probabilities of the number of defective items being equal to 0, 1, 2, ..., 70. Thus,

$$\begin{aligned} (b) \quad P\{\mu \leq 70\} &= \sum_{m=0}^{70} P_n(m) = \\ &= \sum_{m=0}^{70} C_{10,000}^m (0.995)^{10,000-m} (0.005)^m \end{aligned}$$

The above examples demonstrate that a direct computation of the probabilities from formula (1) (and also from formula (1')) come up against formidable technical difficulties; the problem therefore arises of finding simple approximate formulas for the probabilities $P_n(m)$ and also for sums of the form

$$\sum_{m=s}^t P_n(m)$$

for large values of n . These problems will be solved in Secs. 12 and 13. We shall now attempt to establish elementary facts dealing with the behaviour of probabilities $P_n(m)$ for constant n . We begin with a study of $P_n(m)$ as a function of m . It is easy to compute that for $0 \leq m < n$,

$$\frac{P_n(m+1)}{P_n(m)} = \frac{n-m}{m+1} \cdot \frac{p}{q}$$

whence it follows that

$$P_n(m+1) > P_n(m)$$

if $(n-m)p > (m+1)q$, that is, if $np - q > m$;

$$P_n(m+1) = P_n(m)$$

if $m = np - q$, and, finally,

$$P_n(m+1) < P_n(m)$$

if $m > np - q$.

We see that the probability $P_n(m)$ first increases with increasing m and then reaches a maximum, after which it diminishes as m continues to increase. If $np - q$ is an integer, the probability $P_n(m)$

assumes maximal value for two values of m , namely for $m_0=np-q$ and $m'_0=np-q+1=np+p$. But if $np-q$ is not an integer, the maximal value is attained by the probability $P_n(m)$ for $m=\bar{m}_0$, which is equal to the least whole number greater than m_0 . The number \bar{m}_0 is called the *most probable value* of μ . We have seen that if $np-q$ is an integer, then μ has two most probable values: m_0 and $m'_0=m_0+1$.

We note that if $np-q < 0$, then

$$P_n(0) > P_n(1) > \dots > P_n(n)$$

and if $np-q=0$, then

$$P_n(0) = P_n(1) > P_n(2) > \dots > P_n(n)$$

Later on we will see that for large values of n all the probabilities $P_n(m)$ become close to zero, but only for m close to the most probable value of μ are the probabilities $P_n(m)$ noticeably different from zero. We will prove this later on; for the present we illustrate what has been said with a numerical example.

Example 3. Let $n=50$, $p=\frac{1}{3}$.

There are two most probable values: $m_0=np-q=16$ and $m_0+1=17$.

The values of the probabilities $P_n(m)$ are given in Table 5 to four decimal places.

TABLE 5

m	$P_n(m)$	m	$P_n(m)$	m	$P_n(m)$
< 5	0.0000	13	0.0679	23	0.0202
5	0.0001	14	0.0879	24	0.0113
6	0.0004	15	0.1077	25	0.0059
7	0.0012	16	0.1178	26	0.0028
8	0.0033	17	0.1178	27	0.0012
9	0.0077	18	0.1080	28	0.0005
10	0.0157	19	0.0910	29	0.0002
11	0.0287	20	0.0704	30	0.0001
12	0.0470	21	0.0503	> 30	0.0000
		22	0.0332		

Sec. 12. The Local Limit Theorem

When we considered the numerical examples of the previous section, we came to the conclusion that for large values of n and m , calculation of the probabilities $P_n(m)$ from formula (1), Sec. 11, involves considerable difficulties. The necessity arises of having asymptotic formulas that permit calculating these probabilities with a sufficient degree of accuracy. A formula of this kind was first found by DeMoivre in 1730 for the special case of the Bernoulli scheme when $p=q=1/2$, and then was generalized by Laplace to the case of arbitrary p different from 0 and 1.

This formula became known as the *local Laplace theorem*; in order to restore historical justice we shall call it the *local theorem of DeMoivre-Laplace*.

We introduce the notation

$$x = \frac{m - np}{\sqrt{npq}} \quad (1)$$

It is clear that x depends both on n and p and on m .

Local Theorem of DeMoivre-Laplace. *If the probability of occurrence of some event A in n independent trials is constant and is equal to p ($0 < p < 1$), then the probability $P_n(m)$ that in each of the trials event A will occur exactly m times satisfies the relation*

$$\sqrt{npq} P_n(m) : \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \rightarrow 1 \quad (n \rightarrow \infty) \quad (2)$$

uniformly in all m for which x lies in some finite interval.

Proof. The proof that we give is based on Stirling's formula (which is familiar from the course of analysis):

$$s! = \sqrt{2\pi s} \cdot s^s e^{-s} e^{\theta_s}$$

in which the remainder exponent θ_s satisfies the inequality

$$|\theta_s| \leq \frac{1}{12s} \quad (3)$$

We note that formula (1) may be written differently:

$$m = np + x\sqrt{npq} \quad (1')$$

whence it follows that

$$n - m = nq - x\sqrt{npq} \quad (1'')$$

The last two equations permit us to conclude that if x remains bounded by certain constants a and b , then both m and $n - m$ tend to infinity as $n \rightarrow \infty$.

Employment of Stirling's formula gives us

$$P_n(m) = \frac{n!}{m!(n-m)!} p^m q^{n-m} = \sqrt{\frac{n}{2\pi m(n-m)}} \frac{n^n p^m q^{n-m}}{m^m (n-m)^{n-m}} e^\theta \quad (4)$$

where $\theta = \theta_n - \theta_m - \theta_{n-m}$. By virtue of estimate (3) we have

$$|\theta| < \frac{1}{12} \left(\frac{1}{n} + \frac{1}{m} + \frac{1}{n-m} \right)$$

If $a \leq x \leq b$, then the corresponding values of m and $n-m$ satisfy the inequalities

$$\begin{aligned} m &\geq np + a\sqrt{npq} = np \left(1 + a \sqrt{\frac{q}{np}} \right) \\ n-m &\geq nq - b\sqrt{npq} = nq \left(1 - b \sqrt{\frac{p}{nq}} \right) \end{aligned}$$

and, hence, for all indicated values of m and $n-m$ we have the estimate

$$|\theta| < \frac{1}{12n} \left(1 + \frac{1}{p+a\sqrt{\frac{pq}{n}}} + \frac{1}{q-b\sqrt{\frac{pq}{n}}} \right) \quad (5)$$

This shows us that no matter what the interval (a, b) , in this interval the quantity θ tends to zero uniformly in x as $n \rightarrow \infty$, and consequently the factor e^θ under the same conditions uniformly approaches unity.

We now consider the quantity

$$\begin{aligned} \log A_n &= \log \frac{n^n p^m q^{n-m}}{m^m (n-m)^{n-m}} = \log \left(\frac{np}{m} \right)^m + \log \left(\frac{nq}{n-m} \right)^{n-m} = \\ &= -m \log \frac{m}{np} - (n-m) \log \frac{n-m}{nq} = \\ &= -(np + x\sqrt{npq}) \log \left(1 + x \sqrt{\frac{q}{np}} \right) - \\ &\quad -(nq - x\sqrt{npq}) \log \left(1 - x \sqrt{\frac{p}{nq}} \right) \end{aligned}$$

Within the conditions of the theorem, the quantities $x \sqrt{\frac{q}{np}}$ and $x \sqrt{\frac{p}{nq}}$ may be made arbitrarily small for sufficiently large n and so we can take advantage of expanding the functions $\log \left(1 + x \sqrt{\frac{q}{np}} \right)$ and $\log \left(1 - x \sqrt{\frac{p}{nq}} \right)$ in a power series. Con-

fining ourselves to the first two terms, we find

$$\begin{aligned}\log\left(1+x\sqrt{\frac{q}{np}}\right) &= x\sqrt{\frac{q}{np}} - \frac{1}{2}\frac{qx^2}{np} + O\left(\frac{1}{n^{3/2}}\right) \\ \log\left(1-x\sqrt{\frac{p}{nq}}\right) &= -x\sqrt{\frac{p}{nq}} - \frac{1}{2}\frac{px^2}{nq} + O\left(\frac{1}{n^{3/2}}\right)\end{aligned}$$

The estimates of the remainder terms are uniform in any finite interval of variation of x . Thus

$$\begin{aligned}\log A_n &= -(np+x\sqrt{npq})\left[x\sqrt{\frac{q}{np}} - \frac{1}{2}\frac{qx^2}{np} + O\left(\frac{1}{n^{3/2}}\right)\right] - \\ &\quad -(nq-x\sqrt{npq})\left[-x\sqrt{\frac{p}{nq}} - \frac{1}{2}\frac{px^2}{nq} + O\left(\frac{1}{n^{3/2}}\right)\right] = \\ &\quad = -\frac{x^2}{2} + O\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

Hence, the following relation holds uniformly in x in any finite interval $a \leq x \leq b$:

$$A_n : e^{-\frac{x^2}{2}} \rightarrow 1 \quad (6)$$

Further, we have

$$\sqrt{\frac{n}{m(n-m)}} = \frac{1}{\sqrt{npq}} \sqrt{\frac{1}{\left(1+x\sqrt{\frac{q}{np}}\right)\left(1-x\sqrt{\frac{p}{nq}}\right)}} \quad (7)$$

Under the conditions of the theorem, the second factor on the right-hand side of this equation tends to unity as $n \rightarrow \infty$ and does so uniformly in each finite interval of variation of x .

It will readily be seen that relations (5), (6) and (7) prove our theorem.

Now we can complete our computations in the examples of the preceding section.

Example. In Example 2, Sec. 11, we had to determine $P_n(m)$ for $n=10,000$, $m=40$, $p=0.005$. From the theorem just proved we have

$$P_n(m) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{1}{2}\left(\frac{m-np}{\sqrt{npq}}\right)^2}$$

For our example

$$\begin{aligned}\sqrt{npq} &= \sqrt{10,000 \times 0.005 \times 0.995} = \sqrt{49.75} \approx 7.05 \\ \frac{m-np}{\sqrt{npq}} &\approx -1.42\end{aligned}$$

Consequently,

$$P_n(m) \approx \frac{1}{7.05 \sqrt{2\pi}} e^{-\frac{1.42^2}{2}}$$

The function

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is tabulated; an abbreviated table of the values of this function is given at the end of the book (see the Appendix). From this table we find

$$P_n(m) \approx \frac{0.1456}{7.05} \approx 0.0206$$

Exact computations without using the DeMoivre-Laplace theorem give

$$P_n(m) \approx 0.0197$$

To illustrate the nature of the approximations given by the DeMoivre-Laplace theorem and also for a geometric explanation of the analytical transformations carried out in its proof, we shall consider a numerical example.

TABLE 6.
 $n = 4$

m	0	1	2	3	4
$P_n(m)$	0.4096	0.4096	0.1536	0.0256	0.0016
x	-1.00	0.25	1.50	2.75	4.00
$\sqrt{npq} P_n(m)$	0.3277	0.3277	0.1229	0.0205	0.0013
$\varphi(x)$	0.2420	0.3867	0.1295	0.0091	0.0001

Let the probability p be equal to 0.2. Tables 6 to 9 give the values of m , $x = \frac{m - np}{\sqrt{npq}}$, the probabilities $P_n(m)$, the quantities $\sqrt{npq} P_n(m)$, and also the function $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ to the fourth

decimal place for the number of trials, respectively, $n = 4, 25, 100,$ and 400 . In Fig. 8 the ordinates depict the values of the probabilities $P_n(m)$ for various integral values of the abscissa m . It will be seen that $P_n(m)$ uniformly falls off with increasing n .

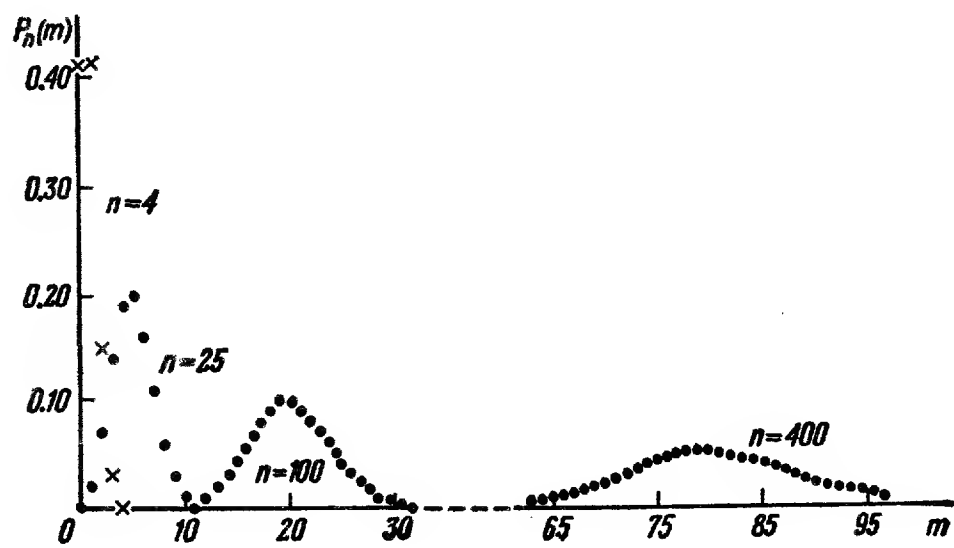


Fig. 8

So that in the figure the points $[m, P_n(m)]$ for the values of n under consideration should not merge with the x -axis, we choose radically different scales for the coordinate axes.

TABLE 7

$n = 25$

m	x	$P_n(m)$	$\sqrt{npq} P_n(m)$	$\varphi(x)$
0	-2.5	0.0037	0.0075	0.0175
1	-2.0	0.0236	0.0472	0.0540
2	-1.5	0.0708	0.1417	0.1295
3	-1.0	0.1358	0.2715	0.2420
4	-0.5	0.1867	0.3734	0.3521
5	0.0	0.1960	0.3920	0.3989
6	0.5	0.1633	0.3267	0.3521
7	1.0	0.1108	0.2217	0.2420
8	1.5	0.0623	0.1247	0.1295
9	2.0	0.0294	0.0589	0.0540
10	2.5	0.0118	0.0236	0.0175
11	3.0	0.0040	0.0080	0.0044
12	3.5	0.0012	0.0023	0.0009
13	4.0	0.0003	0.0006	0.0001
14	4.5	0.0000	0.0000	0.0000
> 14	> 4.5	0.0000	0.0000	0.0000

Consideration of abscissas $x_n = \frac{m-np}{\sqrt{npq}}$ and ordinates $y_n(m) =$

$= \sqrt{npq} P_n(m)$ in place of abscissas m and ordinates $P_n(m)$ signifies:

(1) a translation of the origin to the point $(np, 0)$ located near the abscissa corresponding to the maximal ordinate $P_n(m)$;

(2) an increase in the scale unit along the x -axis by a factor \sqrt{npq} (in other words, compressing the figure along the x -axis \sqrt{npq} times);

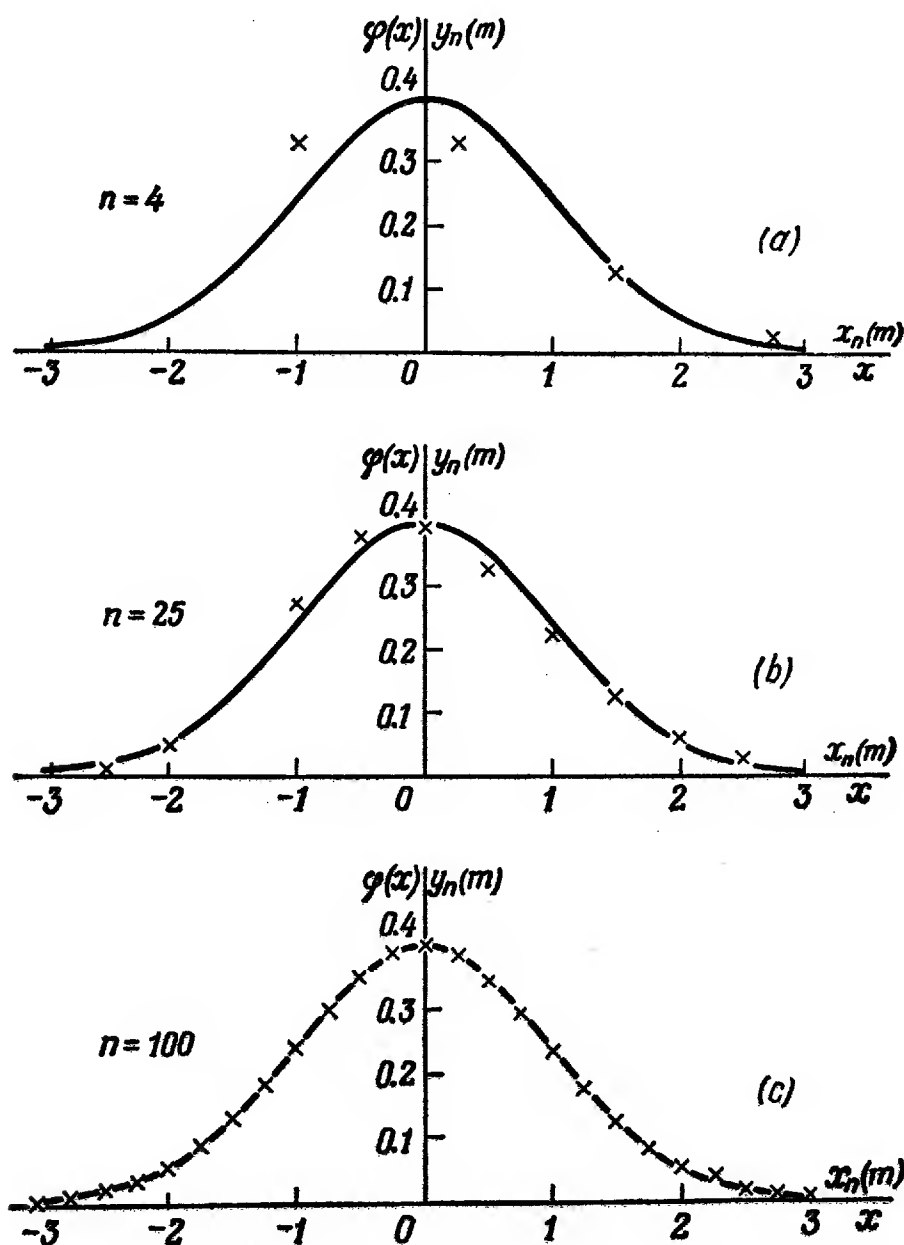


Fig. 9

(3) a decrease in the scale unit along the y -axis by a factor \sqrt{npq} (in other words, expanding the figure along the y -axis \sqrt{npq} times).

Figure 9 (a, b, c) shows: the curve $y = \varphi(x)$ and the points $[m, P_n(m)]$, i.e. the points $[x_n, y_n(m)]$ transformed in the way

just described. We see that already for $n=25$ the points $[x_n, y_n(m)]$ merge in the graph with the corresponding points of the curve $y=\varphi(x)$. This coincidence becomes still better for values of n greater than 25.

TABLE 8
 $n=100$

m	x	$P_n(m)$	$\sqrt{npq} P_n(m)$	$\varphi(x)$
8	-3.00	0.0006	0.0023	0.0044
9	-2.75	0.0015	0.0059	0.0091
10	-2.50	0.0034	0.0134	0.0175
11	-2.25	0.0069	0.0275	0.0317
12	-2.00	0.0127	0.0510	0.0540
13	-1.75	0.0216	0.0863	0.0862
14	-1.50	0.0335	0.1341	0.1295
15	-1.25	0.0481	0.1923	0.1826
16	-1.00	0.0638	0.2553	0.2420
17	-0.75	0.0788	0.3154	0.3011
18	-0.50	0.0909	0.3636	0.3521
19	-0.25	0.0981	0.3923	0.3867
20	0.00	0.0993	0.3972	0.3989
21	0.25	0.0946	0.3783	0.3867
22	0.50	0.0849	0.3396	0.3521
23	0.75	0.0720	0.2879	0.3011
24	1.00	0.0577	0.2309	0.2420
25	1.25	0.0439	0.1755	0.1826
26	1.50	0.0316	0.1266	0.1295
27	1.75	0.0217	0.0867	0.0862
28	2.00	0.0141	0.0565	0.0540
29	2.25	0.0088	0.0351	0.0317
30	2.50	0.0052	0.0208	0.0175
31	2.75	0.0029	0.0117	0.0091
32	3.00	0.0016	0.0063	0.0044

To get a clear-cut idea of the extent to which one can use the asymptotic formula of DeMoivre-Laplace for finite n^* , i.e., to replace the binomial law in determining probabilities $P_n(m)$ by the function $y=\varphi(x)$, we give the following example. For

* Very precise estimates of the remainder term are given in S. N. Bernstein's paper "Returning to the Question of the Accuracy of the Limit Formula of Laplace", *Izv. Akad. nauk S.S.S.R.*, Vol. 7, 1943 (In Russian).

TABLE 9
 $n = 400$

m	x	$P_n(m)$	$\sqrt{npq} P_n(m)$	$\varphi(x)$
56	−3.000	0.0004	0.0034	0.0044
57	−2.875	0.0006	0.0051	0.0064
58	−2.750	0.0009	0.0076	0.0091
59	−2.625	0.0014	0.0104	0.0127
60	−2.500	0.0019	0.0156	0.0175
61	−2.375	0.0027	0.0218	0.0238
62	−2.250	0.0037	0.0298	0.0317
63	−2.125	0.0050	0.0399	0.0417
64	−2.000	0.0066	0.0525	0.0540
65	−1.875	0.0089	0.0679	0.0684
66	−1.750	0.0108	0.0862	0.0862
67	−1.625	0.0134	0.1075	0.1065
68	−1.500	0.0164	0.1316	0.1295
69	−1.375	0.0198	0.1583	0.1550
70	−1.250	0.0234	0.1871	0.1827
71	−1.125	0.0271	0.2175	0.2119
72	−1.000	0.0310	0.2483	0.2420
73	−0.875	0.0349	0.2789	0.2721
74	−0.750	0.0385	0.3081	0.3011
75	−0.625	0.0419	0.3317	0.3282
76	−0.500	0.0447	0.3580	0.3521
77	−0.375	0.0471	0.3766	0.3719
78	−0.250	0.0487	0.3919	0.3867
79	−0.125	0.0497	0.3973	0.3957
80	0.000	0.0498	0.3985	0.3989
81	0.125	0.0492	0.3956	0.3957
82	0.250	0.0478	0.3828	0.3867
83	0.375	0.0458	0.3666	0.3719
84	0.500	0.0432	0.3459	0.3521
85	0.625	0.0402	0.3215	0.3282
86	0.750	0.0368	0.2944	0.3011
87	0.875	0.0332	0.2656	0.2721
88	1.000	0.0295	0.2362	0.2420
89	1.125	0.0259	0.2070	0.2119
90	1.250	0.0223	0.1788	0.1826
91	1.375	0.0190	0.1523	0.1550
92	1.500	0.0160	0.1279	0.1295

Table 9 (continued)

m	x	$P_n(m)$	$\sqrt{npq} P_n(m)$	$\varphi(x)$
93	1.625	0.0132	0.1059	0.1065
94	1.750	0.0108	0.0865	0.0862
95	1.875	0.0087	0.0696	0.0684
96	2.000	0.0069	0.0553	0.0540
97	2.125	0.0054	0.0433	0.0417
98	2.250	0.0042	0.0335	0.0317
99	2.375	0.0032	0.0255	0.0238
100	2.500	0.0024	0.0192	0.0175
101	2.625	0.0018	0.0142	0.0127
102	2.750	0.0013	0.0105	0.0091
103	2.875	0.0009	0.0075	0.0064
104	3.000	0.0008	0.0054	0.0044

the sake of simplicity, consider the case $p=q=\frac{1}{2}$ and take only those n for which it is possible to have $x_{nm}=1$; for instance, these can be $n=25, 100, 400, 1156$. Namely for them $x_{nm}=1$ when $m=15, 55, 210, 595$.

For the sake of brevity, put

$$P_n(m) = P_n$$

and

$$\frac{1}{\sqrt{2\pi npq}} e^{-\frac{x_{nm}^2}{2}} = Q_n$$

for $p=q=\frac{1}{2}$ and $x_{nm}=1$.

According to the local theorem of DeMoivre-Laplace, the ratio $\frac{P_n}{Q_n}$ should tend to unity when $n \rightarrow \infty$. The calculation for the values of n given above yields

TABLE 10

n	P_n	Q_n	$P_n - Q_n$	P_n/Q_n
25	0.09742	0.09679	0.00063	1.0065
100	0.04847	0.04839	0.00008	1.0030
400	0.024207	0.024194	0.000013	1.0004
1156	0.014236	0.014234	0.000002	1.0001

Repeating literally all the reasoning in the proof of the local theorem of DeMoivre-Laplace, we can easily obtain the following multidimensional local theorem. Before formulating it we give the notations:

$$q_i = 1 - p_i \quad (i = 1, 2, \dots, k)$$

$$x_i = \frac{m_i - np_i}{\sqrt{np_i q_i}}$$

The quantity x_i depends not only on i (that is, on p_i) but also on n and m_i ; however, to save space we do not introduce any new indices.

Local Theorem. *If the probabilities p_1, p_2, \dots, p_k of the occurrence, respectively, of events $A_1^{(s)}, A_2^{(s)}, \dots, A_k^{(s)}$ in the s th trial do not depend on the number of the trial and are different from 0 and from 1 ($0 < p_i < 1, i = 1, 2, \dots, k$), then the probability $P_n(m_1, m_2, \dots, m_k)$ that in n independent trials the events $A_i^{(s)}$ ($i = 1, 2, \dots, k$) will occur m_i times ($m_1 + m_2 + \dots + m_k = n$) satisfies the relation **

$$\sqrt{n^{k-1}} P_n(m_1, m_2, \dots, m_k) : \frac{e^{-\frac{1}{2} \sum_{i=1}^k q_i x_i^2}}{(2\pi)^{\frac{k-1}{2}} \sqrt{p_1 p_2 \dots p_k}} \rightarrow 1 \quad (n \rightarrow \infty)$$

uniformly in all m_i ($i = 1, 2, \dots, k$) for which the x_i lie in arbitrary finite intervals $a_i \leq x_i \leq b_i$.

Sec. 13. The Integral Limit Theorem

The local limit theorem just derived will be used to derive another limit relation of probability theory that is called the *integral limit theorem*.

The Integral Theorem of DeMoivre-Laplace. *If μ is the number of occurrences of an event in n independent trials, in each of which the probability of the event is equal to p , and $0 < p < 1$, then the following relation holds uniformly in a and b ($-\infty \leq a \leq b \leq +\infty$)*

* This limiting relation is written in homogeneous coordinates; the quantities x_i are connected by the relation $\sum_{i=1}^k x_i \sqrt{p_i q_i} = 0$, which is readily derived from the relations $\sum m_i = n$ and $\sum p_i = 1$. If we want x_i to be independent variables, one of the arguments, for example x_k , must be eliminated from formula (2).

as $n \rightarrow \infty$:

$$\mathbf{P} \left\{ a \leq \frac{\mu - np}{\sqrt{npq}} < b \right\} - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz \rightarrow 0$$

Proof. For the sake of brevity we introduce the notation

$$P_n(a, b) = \mathbf{P} \left\{ a \leq \frac{\mu - np}{\sqrt{npq}} < b \right\}$$

This probability is obviously equal to the sum $\sum P_n(m)$ extended to those values of m for which $a \leq x_m < b$, where, as before, $x_m = \frac{m - np}{\sqrt{npq}}$.

Now let us define the function $y = \Pi_n(x)$ as follows:

$$y = \Pi_n(x) = \begin{cases} 0 & \text{for } x < x_0 = -\frac{np}{\sqrt{npq}} \\ 0 & \text{for } x \geq x_n + \frac{1}{\sqrt{npq}} = \frac{1 + nq}{\sqrt{npq}} \\ \sqrt{npq} P_n(m) & \text{for } x_m \leq x < x_{m+1} \quad (m = 0, 1, \dots, n) \end{cases}$$

Obviously, the probability $P_n(m)$ is equal to the area bounded by the curve $y = \Pi_n(x)$, the x -axis and the ordinates at points $x = x_m$ and $x = x_{m+1}$, that is,

$$P_n(m) = \sqrt{npq} P_n(m) (x_{m+1} - x_m) = \int_{x_m}^{x_{m+1}} \Pi_n(x) dx$$

Whence it follows that the desired probability $P_n(a, b)$ is equal to the area between the curve $y = \Pi_n(x)$, the x -axis and the ordinates at points $x_{\underline{m}}$ and $x_{\overline{m}}$, where \underline{m} and \overline{m} are defined by the inequalities

$$a \leq x_{\underline{m}} < a + \frac{1}{\sqrt{npq}}, \quad b \leq x_{\overline{m}} < b + \frac{1}{\sqrt{npq}}$$

And so

$$P_n(a, b) = \int_{x_{\underline{m}}}^{x_{\overline{m}}} \Pi_n(x) dx = \int_a^b \Pi_n(x) dx + \int_b^{x_{\overline{m}}} \Pi_n(x) dx - \int_a^{x_{\underline{m}}} \Pi_n(x) dx$$

Since the maximal value of the probability $P_n(m)$ lies at the value $m_0 = [n + 1)p]$, the maximal value of $\Pi_n(x)$ falls in the interval

$$0 \leq \frac{m_0 - np}{\sqrt{npq}} \leq x < \frac{m_0 + 1 - np}{\sqrt{npq}} \leq \frac{2}{\sqrt{npq}}$$

It is in this interval that the local theorem of DeMoivre-Laplace is operative and we can therefore conclude that for all sufficiently large values of n

$$\max \Pi_n(x) < 2 \frac{1}{\sqrt{2\pi}} \max e^{-\frac{x^2}{2}} = \sqrt{\frac{2}{\pi}}$$

From this we first of all draw the conclusion that

$$\begin{aligned} |\rho_n| &= \left| \int_b^{\bar{x}_m} \Pi_n(x) dx - \int_a^{\bar{x}_m} \Pi_n(x) dx \right| \leq \int_b^{\bar{x}_m} \max \Pi_n(x) dx + \\ &+ \int_a^{\bar{x}_m} \max \Pi_n(x) dx < \sqrt{\frac{2}{\pi}} (-b + \bar{x}_m + \bar{x}_m - a) \leq 2 \sqrt{\frac{2}{\pi n p q}} \end{aligned}$$

and that, consequently,

$$\lim_{n \rightarrow \infty} \rho_n = 0$$

Thus, $P_n(a, b)$ differs from $\int_a^b \Pi_n(x) dx$ only by an infinitesimal.

We first assume that a and b are finite numbers. On this assumption, in accord with the local theorem for $a \leq x_m < b$,

$$\Pi_n(x_m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_m^2}{2}} [1 + \alpha_n(x_m)]$$

where $\alpha_n(x_m) \rightarrow 0$ uniformly in x_m as $n \rightarrow \infty$. It is obvious that for the intermediate values of the argument as well,

$$\Pi_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [1 + \alpha_n(x)]$$

and $\lim_{n \rightarrow \infty} \max_{a \leq x < b} \alpha_n(x) = 0$. Indeed, for any m in the interval $x_m \leq x < x_{m+1}$ we have

$$\Pi_n(x) = \Pi_n(x_m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} [1 + \alpha_n(x)]$$

where

$$\alpha_n(x) = e^{\frac{x^2 - x_m^2}{2}} [\alpha_n(x_m) + 1] - 1$$

Since

$$\frac{x^2 - x_m^2}{2} \leq |x| \cdot |x - x_m| < \frac{\max(|a|, |b|)}{\sqrt{npq}}$$

it follows that

$$\lim_{n \rightarrow \infty} \max_{a \leq x < b} \alpha_n(x) = 0$$

Collecting together all the estimates, we obtain

$$P_n(a, b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx + R_n$$

where

$$R_n = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} \alpha_n(x) dx + \rho_n$$

Since

$$|R_n| \leq \max_{a \leq x < b} |\alpha_n(x)| \cdot \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx + \rho_n$$

it is clear from the foregoing that

$$\lim_{n \rightarrow \infty} R_n = 0$$

The theorem is now proved on the particular assumption made in the course of the proof. We now have to get rid of this restriction.

For this purpose, we first of all note that *

$$\frac{1}{\sqrt{2\pi}} \int e^{-\frac{z^2}{2}} dz = 1$$

For this reason, for any $\varepsilon > 0$ it is possible to choose an A so large that

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-A}^A e^{-\frac{z^2}{2}} dz &> 1 - \frac{\varepsilon}{4}, \\ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-A} e^{-\frac{z^2}{2}} dz &= \frac{1}{\sqrt{2\pi}} \int_A^{\infty} e^{-\frac{z^2}{2}} dz < \frac{\varepsilon}{8} \end{aligned}$$

Also, in accordance with what has been proved, choose an n so large that for $-A \leq a \leq b \leq A$ we will have

$$\left| P_n(a, b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz \right| < \frac{\varepsilon}{4}$$

* Here and henceforward, definite integrals without indicated limits are taken from $-\infty$ to $+\infty$.

Then it is obvious that

$$P_n(-A, A) > 1 - \frac{\varepsilon}{2}$$

$$P(-\infty, -A) + P(A, +\infty) = 1 - P(-A, A) < \frac{\varepsilon}{2}$$

Now let us prove that for any a and b ($-\infty \leq a \leq b \leq +\infty$) we will have

$$\left| P_n(a, b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz \right| < \varepsilon$$

thus obviously completing the proof of the Laplace theorem.

To do this we must separately examine different cases of the location of points a and b on the straight line relative to the interval $(-A, A)$. For example, let us take the case $a \leq -A$, $b \geq A$ (the others are left to the reader).

In this case

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz &= \frac{1}{\sqrt{2\pi}} \left(\int_a^{-A} + \int_{-A}^A + \int_A^b e^{-\frac{z^2}{2}} dz \right) \\ P_n(a, b) &= P_n(a, -A) + P_n(-A, A) + P_n(A, b) \end{aligned}$$

Therefore

$$\begin{aligned} \left| P_n(a, b) - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz \right| &\leq \left| P_n(a, -A) - \frac{1}{\sqrt{2\pi}} \int_a^{-A} e^{-\frac{z^2}{2}} dz \right| + \\ &+ \left| P_n(-A, A) - \frac{1}{\sqrt{2\pi}} \int_{-A}^A e^{-\frac{z^2}{2}} dz \right| + \left| P_n(A, b) - \frac{1}{\sqrt{2\pi}} \int_A^b e^{-\frac{z^2}{2}} dz \right| \leq \\ &\leq P_n(-\infty, -A) + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-A} e^{-\frac{z^2}{2}} dz + \\ &+ \left| P_n(-A, A) - \frac{1}{\sqrt{2\pi}} \int_{-A}^A e^{-\frac{z^2}{2}} dz \right| + P_n(A, +\infty) + \\ &+ \frac{1}{\sqrt{2\pi}} \int_A^{+\infty} e^{-\frac{z^2}{2}} dz < \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} + \frac{\varepsilon}{8} = \varepsilon \end{aligned}$$

Let us now formulate the integral limit theorem in the general case of a scheme of a sequence of independent trials. As before, let μ_i ($i = 1, 2, \dots, k$) denote the number of occurrences of events $A_i^{(s)}$ ($s = 1, 2, \dots, n$) in n successive trials. Depending on chance, the num-

bers μ_i may assume only values $0, 1, 2, \dots, n$, and since k outcomes are possible in each trial and these outcomes are mutually exclusive, the following equation must hold:

$$\mu_1 + \mu_2 + \dots + \mu_k = n \quad (1)$$

Let us now regard the quantities $\mu_1, \mu_2, \dots, \mu_k$ as the rectangular coordinates of a point in k -dimensional Euclidean space.

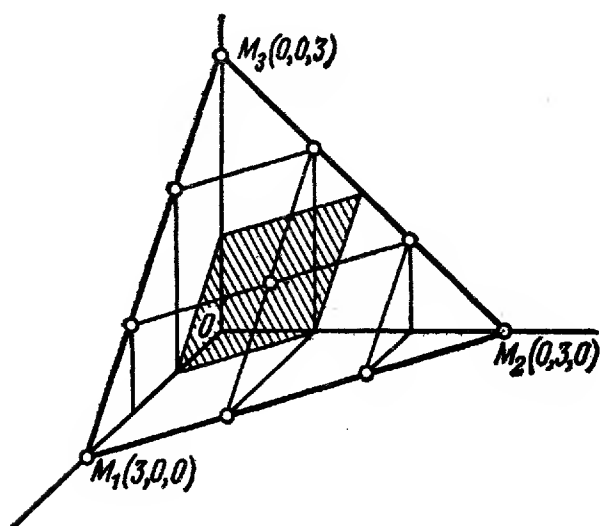


Fig 10

$n=3, k=3$.

Transform the coordinates by means of the formulas

$$x_i = \frac{\mu_i - np_i}{\sqrt{np_i q_i}} \quad (i = 1, 2, \dots, k; q_i = 1 - p_i)$$

In the new coordinates, the equation of the hyperplane (1) will be of the form

$$\sum_{i=1}^k x_i \sqrt{np_i q_i} = 0 \quad (2)$$

We also agree to call *integral* those points of the hyperplane (2) into which the integral points of hyperplane (1) were transformed.

Denote by $P_n(G)$ the probability that as a result of n trials the numbers μ_i ($i = 1, 2, \dots, k$) of occurrences of each of the possible outcomes will be such that the point with coordinates

$$x_i = \frac{\mu_i - np_i}{\sqrt{np_i q_i}}$$

will fall inside the region G .

We then have the following

Theorem. *If in a scheme of a sequence of independent trials there are k possible outcomes in each of the trials, and the probability of each of the outcomes is independent of the number of the trial and is different*

from 0 and from 1, then no matter what the region G of the hyperplane (2), for which the $(k-1)$ -dimensional volume of its boundary is zero, the following relation holds uniformly in G as $n \rightarrow \infty$:

$$P_n(G) \rightarrow \sqrt{\frac{q_1 q_2 \cdots q_k}{(2\pi)^{k-1} \sum_{i=1}^k p_i q_i}} \int_G e^{-\frac{1}{2} \sum_{i=1}^k q_i x_i^2} dv$$

where dv denotes the volume element of the region G and the integral extends over the region G .

We expressed the theorem just formulated in a form in which all the variables x_1, x_2, \dots, x_n play the same role. In the integral theorem of DeMoivre-Laplace, however, we preferred to carry out the reasoning only with the variable $x = x_1$, violating the homogeneity of the variables x_1 and x_2 . Geometrically, this meant that we did not regard the results of the trials themselves (integral points on the straight line $x_1 + x_2 = 0$) but their projections on the x -axis. In similar fashion, by violating the homogeneity in the general case, we can consider integration not over the region G but over its projection G' on some coordinate hyperplane, say, on the plane $x_k = 0$. Volume element dv' in the hyperplane $x_k = 0$ is connected with volume element dv of hyperplane (2) by the relation

$$dv' = dv \cos \varphi$$

where φ is the angle between the indicated hyperplanes. It is easy to calculate that

$$\cos \varphi = \frac{\sqrt{p_k q_k}}{\sqrt{\sum_{i=1}^k p_i q_i}}$$

In the coordinate hyperplane the volume element $dv' = dx_1 dx_2 \dots dx_{k-1}$, we therefore have the equation

$$\begin{aligned} \sqrt{\frac{q_1 q_2 \cdots q_k}{(2\pi)^{k-1} \sum_{i=1}^k p_i q_i}} \int_G e^{-\frac{1}{2} \sum_{i=1}^k q_i x_i^2} dv &= \\ &= \sqrt{\frac{q_1 q_2 \cdots q_{k-1}}{(2\pi)^{k-1} p_k}} \int_{G'} e^{-\frac{1}{2} \sum_{i=1}^k q_i x_i^2} dx_1 \dots dx_{k-1} \end{aligned}$$

In the integrand we must replace x_k by its expression in terms of x_1, x_2, \dots, x_{k-1} :

$$x_k = -\frac{1}{\sqrt{p_k q_k}} \sum_{i=1}^{k-1} \sqrt{p_i q_i} x_i$$

As a result of this substitution we have

$$\sum_{i=1}^k q_i x_i^2 = \sum_{i=1}^{k-1} q_i \left(1 + \frac{p_i}{p_k}\right) x_i^2 + 2 \sum_{1 \leq i < j \leq k-1} x_i x_j \frac{\sqrt{p_i q_i p_j q_j}}{p_k} = Q(x_1, x_2, \dots, x_{k-1}) \quad (3)$$

Thus, the integral limit theorem may be formulated differently:

In the conditions of the integral limit theorem, as $n \rightarrow \infty$,

$$P(G) \rightarrow \sqrt{\frac{q_1 q_2 \dots q_{k-1}}{(2\pi)^{k-1} p_k}} \int_{G'} e^{-\frac{1}{2} Q(x_1, x_2, \dots, x_{k-1})} dx_1 dx_2 \dots dx_{k-1} \quad (4)$$

The integral theorem of DeMoivre-Laplace is a special case of the theorem just proved: it is readily obtainable from formula (4).

To do this, it is sufficient to note that in the Bernoulli scheme $k=2$, $p=p_1$, $q=p_2=1-p$.

For $k=3$ formula (4) takes on the following form:

$$P(G) \rightarrow \sqrt{\frac{q_1 q_2}{(2\pi)^2 p_3}} \int_{G'} e^{-\frac{1}{2} Q(x_1, x_2)} dx_1 dx_2$$

where

$$\begin{aligned} p_3 &= 1 - p_1 - p_2, \\ Q(x_1, x_2) &= q_1 \left(1 + \frac{p_1}{p_3}\right) x_1^2 + q_2 \left(1 + \frac{p_2}{p_3}\right) x_2^2 + 2 \frac{\sqrt{p_1 q_1 p_2 q_2}}{p_3} x_1 x_2 = \\ &= \frac{q_1 q_2}{p_3} \left(x_1^2 + x_2^2 + 2 \sqrt{\frac{p_1 p_2}{q_1 q_2}} x_1 x_2 \right) \end{aligned}$$

A simple calculation shows that

$$p_3 = 1 - p_1 - p_2 = q_1 q_2 - p_1 p_2^*$$

therefore,

$$Q(x_1, x_2) = \frac{1}{1 - \frac{p_1 p_2}{q_1 q_2}} \left(x_1^2 + x_2^2 + 2 \sqrt{\frac{p_1 p_2}{q_1 q_2}} x_1 x_2 \right)$$

Sec. 14. Applications of the Integral Theorem of DeMoivre-Laplace

As a first application of the integral theorem of DeMoivre Laplace, we estimate the probability of the inequality

$$\left| \frac{\mu}{n} - p \right| < \varepsilon$$

* Indeed, since $p_1 + q_1 = 1$ and $p_2 + q_2 = 1$, it follows that $1 - p_1 - p_2 = q_1 - p_2 = q_1(p_2 + q_2) - p_2(p_1 + q_1) = q_1 q_2 - p_1 p_2$.

where $\varepsilon > 0$ is a constant.

We have:

$$\mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| < \varepsilon \right\} = \mathbf{P} \left\{ -\varepsilon \sqrt{\frac{n}{pq}} < \frac{\mu - np}{\sqrt{npq}} < \varepsilon \sqrt{\frac{n}{pq}} \right\}$$

and, consequently, by virtue of the integral theorem of DeMoivre-Laplace

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| < \varepsilon \right\} = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{z^2}{2}} dz = 1$$

And so, no matter what the constant $\varepsilon > 0$, the probability of the inequality $\left| \frac{\mu}{n} - p \right| < \varepsilon$ tends to unity.

This fact was first found by James Bernoulli. It is called the *law of large numbers*, or *Bernoulli's theorem*. Bernoulli's theorem and its numerous generalizations are some of the most important of the theorems of probability theory. It is precisely via them that the theory contacts practice and in them that we see the fundamental success of the application of probability theory to diverse problems of natural science and technology. This will be treated in more detail in the chapter devoted to the law of large numbers, where we will prove the Bernoulli theorem by a simpler method that differs both from the one just given and from Bernoulli's.

We now consider typical problems that lead to the DeMoivre-Laplace theorem.

A total of n independent trials are carried out, for each of which the probability of occurrence of event A is p :

I. What is the probability that the frequency of occurrence of event A will deviate from the probability p by no more than α ? This probability is

$$\begin{aligned} \mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| \leq \alpha \right\} &= \mathbf{P} \left\{ -\alpha \sqrt{\frac{n}{pq}} \leq \frac{\mu - np}{\sqrt{npq}} \leq \alpha \sqrt{\frac{n}{pq}} \right\} \approx \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{-\alpha \sqrt{\frac{n}{pq}}}^{\alpha \sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\alpha \sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}} dx \end{aligned}$$

II. What is the least number of trials that must be carried out so that, with a probability not less than β , the frequency should deviate from the probability by no more than α ? We must determine n from the inequality

$$\mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| \leq \alpha \right\} \geq \beta$$

We replace the probability on the left-hand side of the inequality approximately by an integral using the DeMoivre-Laplace theorem. For a determination of n this yields the inequality

$$\frac{2}{\sqrt{2\pi}} \int_0^{\alpha \sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}} dx \geq \beta$$

III. For a given probability β and the number of trials n , it is required to determine the boundary of possible variations of $\left| \frac{\mu}{n} - p \right|$. In other words, knowing β and n , it is necessary to find α , for which

$$\mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| < \alpha \right\} = \beta$$

For determining α , the integral theorem of Laplace yields the equation

$$\frac{2}{\sqrt{2\pi}} \int_0^{\alpha \sqrt{\frac{n}{pq}}} e^{-\frac{x^2}{2}} dx = \beta$$

The numerical solutions of all the problems we have considered involved evaluating the integral

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz \quad (1)$$

for any values of x and solving the inverse problem: from the value of the integral $\Phi(x)$ compute the appropriate value of the argument x . These calculations require special tables, since for $0 < x < \infty$ the integral (1) is not expressible in closed form in terms of elementary functions. Such tables have been compiled and are given at the end of the book (see the Appendix).

Figure 11 gives a pictorial idea of the function $\Phi(x)$. Using the table of the values of the function $\Phi(x)$ it is also possible to evaluate the integral (using the formula $J(a, b) = \Phi(b) - \Phi(a)$)

$$J(a, b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz$$

The table of the function $\Phi(x)$ is compiled solely for positive x ; for negative x , the function $\Phi(x)$ is found from the equation

$$\Phi(-x) = -\Phi(x)$$

We are now in a position to complete the solution of Example 1 of Sec. 11.

Example 1. In Example 1 of Sec. 11 we had to find the probability

$$P = \sum \mathbf{P} \{ \mu = m \}$$

where the sum is extended to those values of m for which

$$|m - 2.7 \times 10^{22}| \geq 2.7 \times 10^{12}$$

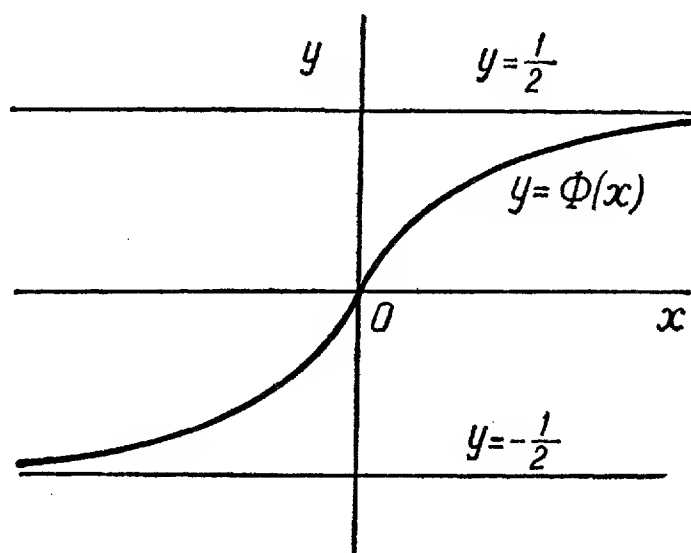


Fig. 11

provided that the total number of trials $n = 5.4 \times 10^{22}$ and $p = \frac{1}{2}$. Since

$$p = \mathbf{P} \left\{ \frac{|\mu - np|}{\sqrt{npq}} \geq \frac{2.7 \times 10^{12}}{\sqrt{5.4 \times 10^{22} \times \frac{1}{4}}} \right\} \approx \mathbf{P} \left\{ \frac{|\mu - np|}{\sqrt{npq}} \geq 2.33 \times 10 \right\}$$

by virtue of the Laplace theorem

$$P \approx \frac{2}{\sqrt{2\pi}} \int_{2.33 \times 10}^{\infty} e^{-\frac{x^2}{2}} dx$$

Since

$$\int_z^{\infty} e^{-\frac{x^2}{2}} dx < \frac{1}{z} \int_z^{\infty} x e^{-\frac{x^2}{2}} dx = \frac{1}{z} e^{-\frac{z^2}{2}}$$

it follows that

$$P < \frac{1}{\sqrt{2\pi} \times 10} e^{-2.7 \times 100} < 10^{-100}$$

To get an idea of just how small this probability is, suppose a sphere of radius 6,000 km is filled with white sand, one grain of which is

black and of size 1 mm³. One grain of sand is taken at random from this mass, what is the probability that it will be black?

It is easy to calculate that the volume of the 6,000-km-radius sphere is slightly less than 10³⁰ mm³ and, consequently, the probability of extracting a black grain of sand is somewhat greater than 10⁻³⁰.

Example 2. In Example 2 of Sec. 11 we had to find the probability that the number of defective items would not exceed seventy if the probability for each item being defective was $p=0.005$ and the number of items was 10,000. From the theorem just proved, this probability is

$$\begin{aligned} \mathbf{P} \{ \mu \leq 70 \} &= \mathbf{P} \left\{ -\frac{50}{\sqrt{49.75}} \leq \frac{\mu - np}{\sqrt{npq}} \leq \frac{20}{\sqrt{49.75}} \right\} = \\ &= \mathbf{P} \left\{ -7.09 \leq \frac{\mu - np}{\sqrt{npq}} \leq 2.84 \right\} \approx \frac{1}{\sqrt{2\pi}} \int_{-7.09}^{2.84} e^{-\frac{z^2}{2}} dz = \\ &= \Phi(2.84) - \Phi(-7.09) = \Phi(2.84) + \Phi(7.09) = 0.9975 \end{aligned}$$

The tables do not contain values of the function $\Phi(x)$ for $x=7.09$, so we replaced it with half that value, thus committing an error less than 10⁻¹⁰.

Naturally, in the examples of this section and the previous one, as in any other problems relating to the determination of probabilities $P_n(m)$ for certain finite values of m and n by the asymptotic DeMoivre-Laplace formulas, it is required to estimate the error due to such a replacement. For a very long time, the DeMoivre-Laplace theorems were applied to the solution of similar problems without a satisfactory estimate of the remainder term. A purely empirical confidence developed that for n of the order of several hundreds or more and for p not too close to 0 or 1, the DeMoivre-Laplace theorems give satisfactory results. At present we have sufficiently good estimates of errors resulting from the use of the asymptotic DeMoivre-Laplace formula.*

Let us also examine the generalization of Bernoulli's theorem to the case of a general scheme of a sequence of independent trials. In each trial let there be k possible outcomes, the probability of each of which is equal, respectively, to p_1, p_2, \dots, p_k and let $\mu_1, \mu_2, \dots, \mu_k$ be the numbers of occurrences of each outcome in the sequence of n independent trials. We determine the probability of a simultaneous realization of the inequalities

$$\left| \frac{\mu_1}{n} - p_1 \right| < \varepsilon_1, \quad \left| \frac{\mu_2}{n} - p_2 \right| < \varepsilon_2, \quad \dots, \quad \left| \frac{\mu_k}{n} - p_k \right| < \varepsilon_k \quad (2)$$

that is, of the inequalities

$$|x_1| < \varepsilon_1 \sqrt{\frac{n}{p_1 q_1}}, \quad |x_2| < \varepsilon_2 \sqrt{\frac{n}{p_2 q_2}}, \quad \dots, \quad |x_k| < \varepsilon_k \sqrt{\frac{n}{p_k q_k}}$$

* See, for example, the paper by S. N. Bernstein cited on p. 81.

Strictly speaking, the last of these inequalities is a corollary to the preceding ones, since in accordance with (2) of Sec. 13, the first $k-1$ of the inequalities (2) yield the estimate

$$|x_k| = \left| -\sum_{i=1}^{k-1} \sqrt{\frac{p_i q_i}{p_k q_k}} x_i \right| \leq \sum_{i=1}^{k-1} \sqrt{\frac{p_i q_i}{p_k q_k}} \varepsilon_i \quad (3)$$

According to (4) of Sec. 13, the probability of the first $k-1$ inequalities (2) and hence also of inequality (3) has as its limit, as $n \rightarrow \infty$, the integral

$$\sqrt{\frac{q_1 q_2 \dots q_{k-1}}{(2\pi)^{k-1} p_k}} \int \dots \int e^{-\frac{1}{2} Q(x_1, \dots, x_k)} dx_1 dx_2 \dots dx_{k-1} = 1$$

Sec. 15. Poisson's Theorem

From the proof of the local theorem of DeMoivre-Laplace it may be noted that the asymptotic representation of the probability $P_n(m)$ by means of the function $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ deteriorates the greater the probability p differs from one half, that is, the smaller the values of p or q that have to be considered; this representation fails for $p=0, q=1$, and also for $p=1, q=0$. However, a substantial range of problems necessitates finding probabilities $P_n(m)$ for just such small values of p .* So that the DeMoivre-Laplace theorem should in that case yield a result with only a slight error it is necessary that the number n of trials be very great. The problem therefore arises of finding an asymptotic formula specially adapted to the case of small p . Such a formula was found by Poisson.

We consider a double sequence of events

$$\begin{array}{ccccccc} E_{11}, & & & & & & \\ E_{21}, & E_{22}, & & & & & \\ E_{31}, & E_{32}, & E_{33}, & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ E_{n1}, & E_{n2}, & E_{n3}, & \dots, & E_{nn}, & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}$$

in which the events of one row are mutually independent and each has a probability p_n that depends solely on the number of the row. By μ_n we denote the number of events that actually occur in the n th row.

* Or for small values of q as well it is obvious, however, that the problems of seeking asymptotic formulas for $P_n(m)$ for small values of p and q reduce to one another.

Poisson's Theorem. If $p_n \rightarrow 0$ as $n \rightarrow \infty$, then

$$\mathbf{P} \{ \mu_n = m \} - \frac{a_n^m}{m!} e^{-a_n} \rightarrow 0 \quad (1)$$

where

$$a_n = np_n$$

Proof. It is obvious that

$$\begin{aligned} P_n(m) &= \mathbf{P} \{ \mu_n = m \} = C_n^m p_n^m (1 - p_n)^{n-m} = \\ &= \frac{n!}{m! (n-m)!} \left(\frac{a_n}{n} \right)^m \left(1 - \frac{a_n}{n} \right)^{n-m} = \\ &= \frac{a_n^m}{m!} \left(1 - \frac{a_n}{n} \right)^n \frac{\left(1 - \frac{1}{n} \right) \left(1 - \frac{2}{n} \right) \dots \left(1 - \frac{m-1}{n} \right)}{\left(1 - \frac{a_n}{n} \right)^m} \quad (2) \end{aligned}$$

Let m be fixed. We choose an arbitrary $\varepsilon > 0$. Then it is possible to choose $A = A(\varepsilon)$ so large that for $a \geq A$ we would have

$$\frac{a^m}{m!} e^{-\frac{1}{2}a} \leq \frac{\varepsilon}{2}$$

Let us first consider those numbers of n for which $a_n \geq A$. For these n , we have, from the inequality $1 - x < e^{-x}$, $0 \leq x \leq 1$:

$$P_n(m) \leq \frac{a_n^m}{m!} e^{-\frac{n-m}{n}a_n} \leq \frac{\varepsilon}{2} \text{ for } n \geq 2m$$

and

$$\frac{a_n^m}{m!} e^{-a_n} < \frac{\varepsilon}{2}$$

Therefore, for the indicated n ,

$$\left| P_n(m) - \frac{a_n^m}{m!} e^{-a_n} \right| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

We now consider those numbers of n for which $a_n \leq A$. Since

$$\lim_{n \rightarrow \infty} \left\{ \left(1 - \frac{a_n}{n} \right)^n - e^{-a_n} \right\} = 0 \text{ for } a_n \leq A \text{ and for constant } m,$$

$$\lim_{n \rightarrow \infty} \frac{\left(1 - \frac{1}{n} \right) \left(1 - \frac{2}{n} \right) \dots \left(1 - \frac{m-1}{n} \right)}{\left(1 - \frac{a}{n} \right)^m} = 1$$

it follows that by virtue of formula (2) for $n \geq n_0(\varepsilon)$

$$\left| P_n(m) - \frac{a_n^m}{m!} e^{-a_n} \right| < \varepsilon$$

and the proof is complete.

We note that the Poisson theorem is also valid when the probability of the event A is zero in each trial. In this case, $a_n = 0$.

We denote

$$P(m) = \frac{a^m}{m!} e^{-a}$$

The probability distribution thus obtained is called *Poisson's law* or the *Poisson distribution*.

It is easy to calculate that the quantities $P(m)$ satisfy the equality $\sum_m P(m) = 1$. Let us study the behaviour of $P(m)$ as a function of m . With this purpose in mind, consider the relation

$$\frac{P(m)}{P(m-1)} = \frac{a}{m}$$

We see that if $m > a$, then $P(m) < P(m-1)$, but if $m < a$, then $P(m) > P(m-1)$; if, finally, $m = a$, then $P(m) = P(m-1)$. From this we conclude that the quantity $P(m)$ increases with increasing m from 0 to $m_0 = [a]$ and falls off with further increases in m . If a is an integer, then $P(m)$ has two maximal values: at $m_0 = a$ and at $m'_0 = a - 1$.

The following are examples.

Example 1. The probability of hitting a target is 0.001 for each shot. Find the probability of hitting the target with two or more bullets if the number of shots is 5,000.*

Taking each shot as a trial and hitting the target as an event, we can take advantage of Poisson's theorem to compute the probability $P\{\mu_n \geq 2\}$. In our case,

$$a_n = np = 0.001 \times 5,000 = 5$$

The probability sought for is

$$P\{\mu_n \geq 2\} = \sum_{m=2}^{5,000} P_n(m) = 1 - P_n(0) - P_n(1)$$

By the Poisson theorem

$$P_n(0) \approx e^{-5}, \quad P_n(1) \approx 5e^{-5}$$

Therefore

$$P\{\mu_n \geq 2\} \approx 1 - 6e^{-5} \approx 0.9596$$

The probability $P_n(m)$ takes on a maximal value for $m=4$ and $m=5$. To four decimal places, these probabilities are equal to

$$P(4) = P(5) \approx 0.1751$$

* In the Great Patriotic War, the conditions of our problem were realized in small-arms fire against aircraft. An aircraft can be shot down only if hit in a vulnerable spot: motor, pilot, fuel tank, etc. The probability of hitting these

Using an exact formula, the computations yield (to the fourth decimal place) $P_{5,000}(0)=0.0067$, $P_{5,000}(1)=0.0336$ and, consequently,

$$P\{\mu_n \geq 2\} = 0.9597$$

The error due to use of the asymptotic formula is less than 0.01% of the value being computed.

Example 2. A worker at a spinning mill attends several hundred spindles, each of which spins its own skein of yarn. In the process of winding, the yarn breaks due to nonuniformity of tension, unevenness and other causes at chance instants of time. It is important to know how frequently such breaks can occur under one or another set of conditions (grade of yarn, speed of spindles, etc.).

Assuming that a worker attends 800 spindles and the probability of yarn breakage on each spindle during a certain interval of time τ is 0.005, find the most probable number of breaks and the probability that during the time interval τ there will be no more than 10 breaks.

Since

$$a_n = np = 0.005 \times 800 = 4$$

there will be two most probable numbers of breaks in the time interval τ : 3 and 4. Their probabilities are

$$P_{800}(3) = P_{800}(4) = C_{800}^4 \times 0.005^4 \times 0.995^{796}$$

Using Poisson's formula we have

$$P_{800}(3) = P_{800}(4) \approx \frac{4^3}{3!} e^{-4} = \frac{32}{3} \times e^{-4} = 0.1954$$

The exact value of $P_{800}(3) = P_{800}(4) = 0.1945$. The probability that the number of breaks in a time interval τ will not exceed 10 is equal to

$$P\{\mu_n \leq 10\} = \sum_{m=0}^{10} P_{800}(m) = 1 - \sum_{m=11}^{\infty} P_{800}(m)$$

By virtue of the Poisson theorem,

$$P_{800}(m) \approx \frac{4^m}{m!} e^{-4} \quad (m = 0, 1, 2, \dots)$$

and so

$$P\{\mu_n \leq 10\} = 1 - \sum_{m=11}^{\infty} \frac{4^m}{m!} e^{-4}$$

vulnerable spots with a single shot is extremely small, but, as a rule, a whole unit fired at once and the total number of shots was considerable. The probability of one or two bullets making a successful strike was then rather appreciable. This was also found to be so in actual cases.

But

$$\sum_{m=11}^{\infty} \frac{4^m}{m!} e^{-4} > \left(\frac{4^{11}}{11!} + \frac{4^{12}}{12!} + \frac{4^{13}}{13!} \right) e^{-4} = \frac{4^{12} \cdot 14}{11! 139} e^{-4} = 0.00276$$

On the other hand,

$$\begin{aligned} \sum_{m=11}^{\infty} \frac{4^m}{m!} e^{-4} &< \frac{4^{11}}{11!} e^{-4} + \frac{4^{12}}{12!} e^{-4} + \frac{4^{13}}{13!} e^{-4} \left[1 + \frac{4}{14} + \left(\frac{4}{14} \right)^2 + \dots \right] = \\ &= \frac{4^{12} \cdot 24}{11! 65} e^{-4} = 0.00284 \end{aligned}$$

Thus,

$$0.99716 \leq \mathbf{P} \{ \mu_n \leq 10 \} \leq 0.99724$$

Just as in the case when applying the DeMoivre-Laplace theorem, we have to estimate the error that results from replacing the exact formula for computing $P_n(m)$ by the asymptotic formula of Poisson.

From the equation

$$\begin{aligned} P_n(0) &= \left(1 - \frac{a_n}{n} \right)^n = e^{n \ln \left(1 - \frac{a_n}{n} \right)} = \\ &= \exp \left\{ -n \sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{a_n}{n} \right)^k \right\} = e^{-a_n} (1 - R_n) \end{aligned}$$

where

$$R_n = 1 - \exp \left\{ -n \sum_{k=2}^{\infty} \frac{1}{k} \left(\frac{a_n}{n} \right)^k \right\}$$

we can readily find this estimate for the case $m=0$. Indeed, since for arbitrary positive x

$$0 < 1 - e^{-x} < x$$

it follows that, no matter what the a_n and n were,

$$0 < R_n < n \sum_{k=2}^{\infty} \frac{1}{k} \left(\frac{a_n}{n} \right)^k$$

Since

$$\begin{aligned} \sum_{k=2}^{\infty} \frac{1}{k} \left(\frac{a_n}{n} \right)^k &\leq \frac{a_n^2}{2n^2} + \frac{1}{3} \sum_{k=3}^{\infty} \left(\frac{a_n}{n} \right)^k = \\ &= \frac{a_n^2}{2n^2} + \frac{a_n^3}{3n^3 \left(1 - \frac{a_n}{n} \right)} = \frac{a_n^2}{6n^2} \cdot \frac{3n - a_n}{n - a_n} < \frac{a_n^2}{2n(n - a_n)} \end{aligned}$$

it follows that

$$0 < R_n < \frac{a_n^2}{2(n-a_n)}$$

From the fact that R_n is nonnegative, we conclude that when $P_n(0)$ is replaced by e^{-a_n} we increase somewhat the probability $P_n(0)$.

Sec. 16. An Illustration of the Scheme of Independent Trials

By way of illustrating the use of the foregoing results in the natural sciences, we consider very schematically the problem of a random walk of a particle on a straight line. This problem may be regarded as the prototype of actual physical problems in the theory of diffusion, Brownian motion, and so forth.

Imagine that at specific instants of time a particle, starting from the position $x=0$, experiences random impacts that displace it to the right or the left one unit of distance. Thus, each time, the particle is shifted one unit to the right or one to the left with a probability of $\frac{1}{2}$. As a result of n impacts the particle will have been displaced to a distance μ . In this problem we clearly have to do with the Bernoulli scheme in its pure form. It then follows that for each n and m we can calculate the probability that $\mu=m$; namely,

$$\mathbf{P} \{ \mu = m \} = \begin{cases} C_n^{\frac{m+n}{2}} \left(\frac{1}{2} \right)^n & \text{if } -n \leq m \leq n \\ 0 & \text{if } |m| > n \end{cases}$$

For large values of n , as follows from the local theorem of DeMoivre-Laplace,

$$\mathbf{P} \{ \mu = m \} \approx \frac{\sqrt{2}}{\sqrt{\pi n}} e^{-\frac{m^2}{2n}} \quad (1)$$

We may regard this formula as follows. Suppose at an initial time there are a large number of particles with coordinate $x=0$. All these particles begin to move along a straight line independently of one another as a consequence of random impacts. Then, after n impacts, the portion of particles that has covered a distance m is given by formula (1).

We of course consider idealized conditions of particle motion; actual molecules move under much more complicated conditions, but the overall result yields a correct *qualitative* picture of the phenomenon.

In physics, more involved examples of random walks are considered. We confine ourselves to just as schematic a consideration of the effects of: (1) a reflecting barrier; (2) an absorbing barrier.

Imagine that at a distance of s units to the right of point $x=0$ there is a reflecting barrier, such that a particle which at some time

hits the barrier is, upon the next impact, returned with probability one in the same direction from which it arrived.

Figure 12 gives the reader a more pictorial view of a particle on a plane (x, t) . The path of the particle will be depicted as a broken line. Each impact advances the particle one unit "upwards" and one unit to the right or the left (with a probability one half each time that $x < s$). Now if $x = s$, then on the next impact the particle will be shifted one unit to the left.

To compute the probability $P\{\mu = m\}$ we do as follows: we mentally eliminate the barrier and allow the particle to move freely as if there were no barrier at all. Figure 12 shows such idealized paths that lead to points A and A' , which are symmetric about the barrier. For an actual particle, moving with reflections, to reach A it is necessary and sufficient for the particle moving in the idealized situation (without the reflecting barrier) to reach either A or A' . But the probability of getting to point A in the idealized situation is obviously equal to

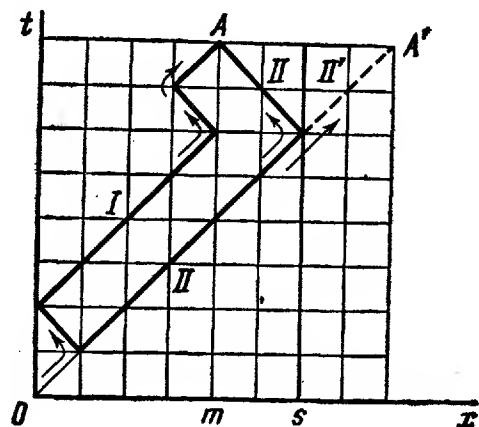


Fig. 12

$$P\{\mu = m\} = \frac{n!}{\left(\frac{m+n}{2}\right)! \left(\frac{n-m}{2}\right)!} \left(\frac{1}{2}\right)^n$$

In exactly the same way, the probability of getting to point A' is equal (the abscissa of A' is $2s - m$) to

$$P\{\mu = 2s - m\} = \frac{n!}{\left(s + \frac{n-m}{2}\right)! \left(\frac{n+m}{2} - s\right)!} \left(\frac{1}{2}\right)^n$$

The desired probability is thus

$$P_n(m; s) = P\{\mu = m\} + P\{\mu = 2s - m\}$$

Taking advantage of the local limit theorem of DeMoivre-Laplace,

$$P_n(m; s) \approx \frac{\sqrt{2}}{\sqrt{\pi n}} \left\{ e^{-\frac{m^2}{2n}} + e^{-\frac{(2s-m)^2}{2n}} \right\}$$

This is the famous formula of Brownian motion theory. It takes on a more symmetric form if the origin is placed at point $x = s$, and, hence, if we pass to a new coordinate z by the formula $z = x - s$. This substitution gives us

$$P_n(z=k) = P_n\{k+s, s\} \approx \frac{\sqrt{2}}{\sqrt{\pi n}} \left\{ e^{-\frac{(k+s)^2}{2n}} + e^{-\frac{(k-s)^2}{2n}} \right\}$$

We now consider the third schematic problem: an absorbing barrier is placed in the path of the particle at point $x=s$. A particle striking this barrier drops out of the motion. Obviously, in this example the probability of getting to point $x=m$ ($m < s$) after n impacts will be less than $P_n(m)$ (that is, less than the probability of getting to this point without the absorbing barrier); denote the desired probability by the symbol $\bar{P}_n(m; s)$.

To compute the probability $\bar{P}_n(m; s)$ we again eliminate mentally the absorbing barrier and allow the particle to move freely along the straight line. If at some time the particle reaches $x=s$, at subsequent instants of time it will go to the right and left of the line $x=s$ (Fig. 13) with the same probability. In exactly the same way, after getting to the straight line $x=s$, the particle can reach both point $A(m, n)$ and point $A'(2s-m, n)$ with the same probability. But the particle can reach A' only after first having reached the position $x=s$; therefore, for any pathway leading to A' there is a path symmetric about the straight line $x=s$ and leading to A ; in exactly the same way, for every prohibited pathway (in actual motion)

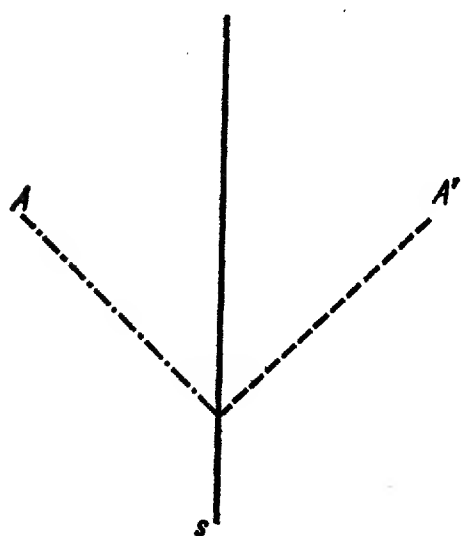


Fig. 13

leading to A there is a path symmetric about $x=s$ that leads to point A' . It will be noted that here we consider the symmetry of pathways only after hitting the straight line $x=s$. The foregoing reasoning shows that when counting the number of favourable cases in actual motion we must eliminate from the paths leading to A in idealized motion the exact number of paths that lead to point A' . It obviously follows from this that

$$\bar{P}_n(m; s) = P\{\mu = m\} - P\{\mu = 2s - m\}$$

By virtue of the local theorem of DeMoivre-Laplace we have

$$\bar{P}_n(m; s) \approx \frac{\sqrt{2}}{\sqrt{\pi n}} \left\{ e^{-\frac{m^2}{2n}} - e^{-\frac{(2s-m)^2}{2n}} \right\}$$

EXERCISES

1. A workman operates 12 machines of the same type. The probability that one machine will require his attention during a time interval of duration τ is $1/3$. What is the probability that:

- (a) during time τ 4 machines will demand the attention of the workman;
 (b) the number of such demands during time τ will lie between 3 and 6 (including the boundaries)?

2. A certain family has 10 children. Considering the probability of birth of a boy and a girl equal to $1/2$, find the probability that in this family

- (a) there are 5 boys and 5 girls;
 (b) the number of boys lies between 3 and 8.

3. In a gathering of 4 persons, the birthdays of three come in one month and that of the fourth in one of the remaining eleven months. Considering the probability of birth of each person in each month equal to $1/12$, find the probability that

- (a) the three persons were born in January and the fourth in October;
 (b) the three were born in some one month and the fourth in one of the other eleven months.

4. In 14,400 tosses of a coin, heads fell 7,428 times. How probable is such a large or larger deviation of the number of heads from np if the coin is symmetric (that is, the probability of throwing heads in each trial is $1/2$)?

5. A total of n devices, each with a power consumption of a kilowatts, are connected to an electric network. At a given time each is consuming power with a probability p . Find the probability that the power consumed at the given time

- (a) will be less than nap ;
 (b) will exceed $rnep$ ($r > 0$) provided that np is great.

6. An educational institution has a student body of 730. The probability that the birthday of a randomly selected student will fall on a definite day of the year is $1/365$ for each of the 365 days. Find:

- (a) the most probable number of students born on January 1;
 (b) the probability that there will be three students with the same birthday.

7. It is known that the probability of producing a drill bit of extra-high brittleness (defective) is 0.02. The bits are packed in boxes of a hundred each. What is the probability that

- (a) a box will have no defective bits;
 (b) the number of defective bits will be less than 3.

How many bits have to be put in a box so that there should be at least 100 good bits with a probability not less than 0.9?

Hint. Take advantage of the Poisson distribution.

8. An insurance company has issued policies to 10,000 persons of the same age and the same social group. The probability of death during the year for each person is 0.006. On January 1 each insured person deposits 12 rubles on his policy and if he dies his beneficiaries receive 1,000 rubles from the company. What is the probability that:

- (a) the company will suffer a loss;
 (b) the company will make a profit of at least 40,000; 60,000; 80,000 rubles?

9. Prove the following theorem: if P and P' are the probabilities of the most probable number of occurrences of an event A in n and $n+1$ independent trials (in each of the trials $P(A)=p$), then $P' \leq P$. The equality is excluded if $(n+1)p$ is not an integer.

10. In the Bernoulli scheme, $p=1/2$. Prove that

$$(a) \quad \frac{1}{2\sqrt{n}} \leq P_{2n}(n) \leq \frac{1}{\sqrt{2n+1}}$$

$$(b) \quad \lim_{n \rightarrow \infty} \frac{P_{2n}(n \pm h)}{P_{2n}(n)} = e^{-z^2}$$

if $\frac{h}{\sqrt{n}} = z$ ($0 \leq z < +\infty$).

11. Prove that for $npq \geq 25$

$$P_n(m) = \frac{1}{\sqrt{2npq}} e^{-\frac{z^2}{2}} \left[1 + \frac{(q-p)(z^3-3z)}{6\sqrt{npq}} \right] + \Delta$$

where

$$z = \frac{m-np}{\sqrt{npq}}, \quad |\Delta| < \frac{0.15 + 0.25|p-q|}{\sqrt{(npq)^3}} |z| e^{-\frac{3}{2}\sqrt{npq}}$$

12. A total of n independent trials have been performed. The probability of the occurrence of event A in the i th trial is p_i ; $P_n(m)$ is the probability of the m -fold occurrence of event A in n trials. Prove that

$$(a) \quad \frac{P_n(1)}{P_n(0)} \geq \frac{P_n(2)}{P_n(1)} \geq \dots \geq \frac{P_n(n)}{P_n(n-1)}$$

(b) $P_n(m)$ first increases and then decreases (if $P_n(0)$ or $P_n(n)$ are not themselves maximal).

13. Prove that for $x > 0$ the function $\int_x^\infty e^{-\frac{z^2}{2}} dz$ satisfies the inequalities

$$\frac{x}{1+x^2} e^{-\frac{1}{2}x^2} \leq \int_x^\infty e^{-\frac{1}{2}z^2} dz \leq \frac{1}{x} e^{-\frac{1}{2}x^2}$$

14. **Banach's match box problem.** A certain mathematician always carries two boxes of matches with him. Whenever he wants a match, he selects one of the boxes at random. Find the probability that when the mathematician draws an empty box, the other box will contain r matches ($r=0, 1, 2, \dots, n$; n is the number of matches initially contained in each box).

15. A total of n machines are connected to an electric transmission line. The probability that a machine consuming power at time t will cease to consume up to time $t+\Delta t$ is equal to $\alpha\Delta t + o(\Delta t)$. If at time t a machine is not consuming any power, then the probability that it will begin consuming prior to time $t+\Delta t$ is equal to $\beta\Delta t + o(\Delta t)$, irrespective of the operation of the other machines. Form differential equations that are satisfied by the probabilities $P_r(t)$ that at time t a total of r machines will be consuming power.

Note. It is easy to indicate the concrete conditions of this problem: the movement of trams, electric welding, power consumption by machine tools with automatic cutoff, and so forth.

16. One workman operates n automatic machines of the same type. If at time t a machine is operating, then the probability that it will require attention prior to time $t+\Delta t$ is equal to $\alpha\Delta t + o(\Delta t)$. If at time t the operator is attending some machine, then the probability that he will complete his job prior to time $t+\Delta t$ is equal to $\beta\Delta t + o(\Delta t)$. Form differential equations satisfied by the probabilities $P_r(t)$ that at time t , $n-r$ machines will be in operation; one is being attended and $r-1$ are in line waiting to be serviced ($P_0(t)$ is the probability that all the machines are in operation).

Note. It is easy to form differential equations in similar fashion for the more complicated problem when N machines are attended by a team of k workmen. It is important for practical purposes to compare the economy of one or another system of organizing the labour. For this purpose, it is necessary to study the steady-state regime, that is, to consider the probabilities $P_r(t)$ as $t \rightarrow \infty$.

It turns out that the work of a team attending kn machines has advantages over one operator attending n machines both in the meaning of better utilization of the operating time of the machine and the working time of the operator.

CHAPTER 3

Markov Chains

Sec. 17. Markov Chains Defined. Transition Matrix

A direct generalization of the scheme of independent trials is a scheme of what are known as *Markov chains*, which were studied systematically for the first time by the noted Russian mathematician A. A. Markov. We will confine ourselves to the fundamentals of his theory.

Imagine that we have a sequence of trials in each of which one and only one of k mutually exclusive events $A_1^{(s)}, A_2^{(s)}, \dots, A_k^{(s)}$ (as in Chapter 2, the superscript denotes the number of the trial) can occur. We say that the sequence of trials forms a *Markov chain*, or more precisely a *simple Markov chain*, if the *conditional probability that event $A_i^{(s+1)}$ ($i=1, 2, \dots, k$) will occur in the $(s+1)$ st trial ($s=1, 2, 3, \dots$), after a known event has occurred in the s th trial, depends solely on the event that occurred in the s th trial and is not modified by supplementary information about the events that occurred in earlier trials.*

A different terminology is frequently employed in stating the theory of Markov chains, and one speaks of a certain physical system S , which at each instant of time can be in one of the states A_1, A_2, \dots, A_k and alters its state only at times $t_1, t_2, \dots, t_n, \dots$. For Markov chains, the probability of passing to some state A_i ($i=1, 2, \dots, k$) at time τ ($t_s < \tau < t_{s+1}$) depends only on the state the system was in at time t ($t_{s-1} < t < t_s$) and does not change if we learn what its states were at earlier times.

By way of illustration we consider two schematic cases.

Example 1. Imagine that a particle located on a straight line moves along the line via random impacts occurring at times t_1, t_2, t_3, \dots . The particle can be at points with integral coordinates $a, a+1, a+2, \dots, b$; at points a and b there are reflecting barriers. Each impact displaces the particle to the right with probability p and to the left with probability $q=1-p$ so long as the particle is not located at a barrier. If the particle is at a barrier, any impact will transfer it one

unit inside the gap between the barriers. We see that this instance of a particle walk is a typical Markov chain. We could just as easily consider the case when the particle is sticking to one of the barriers or to both of them.

Example 2. In Bohr's model of the hydrogen atom, the electron can be in one of the allowed orbits. Denote by A_i the event that the electron lies in the i th orbit. Further assume that changes in the state of the atom can occur only at times t_1, t_2, t_3, \dots (actually, these times are random quantities). The probability of transition from the i th orbit to the j th at time t_s depends only on i and j (the difference $j-i$ depends on the amount of energy by which the charge of the atom changed at time t_s) and does not depend on the orbits the electron occupied in the past.

This case is a Markov chain with an infinite (true, only in principle) number of states; this instance would be incomparably closer to a real situation if the times of transitions of our system to a new state varied continuously.

* * *

We confine ourselves to the statement of the most elementary facts for *homogeneous Markov chains* in which the conditional probability of the occurrence of an event $A_j^{(s+1)}$ in the $(s+1)$ st trial, provided that in the s th trial the event $A_i^{(s)}$ occurred, does not depend on the number of the trial. We call this probability the *transition probability* and denote it by p_{ij} ; in this notation, the first subscript always denotes the result of the previous trial, and the second indicates the state into which the system passes in the subsequent instant of time.

The total probabilistic picture of possible changes that occur during a transition from one trial to the immediately following one is given by the matrix

$$\pi_1 = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdot & \cdot & \cdot & \cdot \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

compiled of the transition probabilities; we will call this matrix the *transition matrix (matrix of transition probabilities)*.

The following examples will serve as illustrations.

Example 3. The system S that we are studying can be in the states A_1, A_2, A_3 ; transition from state to state occurs in accordance with the scheme of a homogeneous Markov chain; the transition probabi-

lities are given by the matrix

$$\pi_1 = \begin{pmatrix} 1/2 & 1/6 & 1/3 \\ 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$$

We see that if the system was in the state A_1 , then after a change of the state by one step it will remain in the same state with a probability of $1/2$, and it will pass to state A_2 with a probability of $1/6$, and to state A_3 with a probability of $1/3$. But if the system was in the state A_2 , then after the transition it can (with equal probability) find itself only in states A_1 and A_3 ; it cannot pass from state A_2 into A_2 . The last row of the matrix shows us that from the state A_3 the system can pass to any one of the possible states with one and the same probability $1/3$.

Example 4. Let us write the transition matrix for the case, described in the first example, of a particle in a random walk between two reflecting barriers. If we denote by A_1 the event consisting in the particle being at a point with coordinate a , by A_2 , it being at a point with coordinate $a+1$, ..., by A_s ($s=b-a+1$), it being at a point with coordinate b , then the transition matrix will be as follows:

$$\pi_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ q & 0 & p & 0 & \dots & 0 \\ 0 & q & 0 & p & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Example 5. We also write the transition matrix for a particle in a random walk between two absorbing barriers. The notations and the conditions remain the same as in Example 4, the only difference being that the particle which passes to state A_1 or A_s remains in those states with a probability of 1:

$$\pi_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ q & 0 & p & 0 & \dots & 0 \\ 0 & q & 0 & p & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Let us point out what conditions have to be satisfied by the elements of the transition matrix. First of all, being probabilities, they must be nonnegative numbers, i.e., for all i and j

$$0 \leq p_{ij} \leq 1$$

Also, from the fact that in the transition from state $A_i^{(s)}$ prior to the $(s+1)$ st trial the system must definitely pass to one and

only one of the states $A_j^{(s+1)}$ after the $(s+1)$ st trial there follows the equation

$$\sum_{j=1}^k p_{ij} = 1 \quad (i = 1, 2, \dots, k)$$

Thus, the sum of the elements of each row of the transition matrix is equal to unity.

Our first problem in the theory of Markov chains consists in determining the transition probability from state $A_i^{(s)}$ in the s th trial to the state $A_j^{(s+n)}$ after n trials. We denote this probability by the symbol $P_{ij}(n)$.

Let us examine some intermediate trial with the number $s+m$. Some one of the possible events $A_r^{(s+m)}$ ($1 \leq r \leq k$) will occur in this trial. In accord with the notations just introduced, the probability of such a transition is equal to $P_{ir}(m)$. And the probability of transition from state $A_r^{(s+m)}$ to state $A_j^{(s+n)}$ is $P_{rj}(n-m)$. By the formula of the total probability,

$$P_{ij}(n) = \sum_{r=1}^k P_{ir}(m) \cdot P_{rj}(n-m) \quad (1)$$

We denote by π_n the transition matrix after n trials:

$$\pi_n = \begin{pmatrix} P_{11}(n) & P_{12}(n) & \dots & P_{1k}(n) \\ \cdot & \cdot & \cdot & \cdot \\ P_{k1}(n) & P_{k2}(n) & \dots & P_{kk}(n) \end{pmatrix}$$

According to (1), the following relation holds between the matrices π_s with different subscripts:

$$\pi_n = \pi_m \cdot \pi_{n-m} \quad (0 < m < n)$$

In particular, for $n=2$, we find:

$$\pi_2 = \pi_1 \cdot \pi_1 = \pi_1^2$$

for $n=3$

$$\pi_3 = \pi_1 \cdot \pi_2 = \pi_2 \cdot \pi_1 = \pi_1^3$$

and, generally, for any n ,

$$\pi_n = \pi_1^n$$

We note a special case of formula (1): for $m=1$

$$P_{ij}(n) = \sum_{r=1}^k p_{ir} P_{rj}(n-1)$$

Example 6. A simple count shows that the two-step transition matrices of Examples 4 and 5 of this section are of the following form:

for a random walk of a particle between reflecting barriers ($s \geq 5$)

$$\begin{pmatrix} q & 0 & p & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & q+pq & 0 & p^2 & 0 & \dots & 0 & 0 & 0 \\ q^2 & 0 & 2pq & 0 & p^2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & \dots & 0 & 0 & 0 & \dots & q & 0 & p \end{pmatrix}$$

for a particle in a random walk between absorbing barriers

$$\begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots \\ q & p & q & 0 & p^2 & \dots & \dots & \dots \\ q^2 & 0 & 2pq & 0 & p^2 & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 1 \end{pmatrix}$$

It is intuitively clear that in the case of reflecting barriers a particle will, after a large number of steps, be able to reach any point between the barriers. But in the case of absorbing barriers, the larger the number of steps a system has covered, the greater the probability that the particle will be absorbed by the barriers.

Sec. 18. Classification of Possible States

The classification of states offered here was described at almost the same time by A. N. Kolmogorov for Markov chains with a countable set of states and by W. Doeblin for Markov chains with a finite set of states.

The state A_i is called *unessential* (or *transient*) if there exist A_j and n such that $P_{ij}(n) > 0$, but $P_{ji}(m) = 0$ for all m . Thus, an unessential state possesses the property that it is possible, with positive probability, to pass from it into other state, but it is no longer possible to return from that state to the original (unessential) state. Of the examples of the preceding section, we consider the fifth: the random walk of a particle between two absorbing barriers. It is easy to see that in this example all the states, except A_1 and A_s , are unessential. Indeed, no matter what the state (different from A_1 and A_s) a particle is in, it can reach both A_1 and A_s with positive probabilities via a finite number of steps, but it cannot return from these states into any other state.

All states not unessential are called *essential*. From the definition it follows that if the states A_i and A_j are essential, then there exist positive m and n such that along with the inequality $P_{ij}(m) > 0$ the inequality $P_{ji}(n) > 0$ also holds. If A_i and A_j are such that for both of them these inequalities hold, given certain m and n , then they are called *communicating* (they are said to communicate). It is clear that if A_i communicates with A_j , and A_j communicates with A_k , then A_i also communicates with A_k . Thus, all essential states can be

partitioned into *classes* such that all states belonging to a single class communicate and those belonging to different classes do not communicate. All states in Examples 3 and 4 of the preceding section are essential and in each case form a unique class of states.

Since for the essential state A_i and the unessential state A_j the equation $P_{ij}(m)=0$ holds for any m , we can draw the following conclusion: if a system has reached one of the states of a definite class of essential states, it can no longer leave that class. Example 5 exhibits two classes of essential states, each of which consists of a single element: one class is the state A_1 and the other is the state A_8 .

Let us now examine more closely the mechanism of transition from state to state inside one class. To do this, take some essential state A_i and denote by M_i the set of all integers m for which $P_{ii}(m) > 0$. This set cannot be empty by virtue of the definition of an essential state. It is immediately obvious that if the numbers m and n are contained in the set M_i , then their sum, $m+n$, also belongs to this set. Denote by d_i the greatest common divisor of all the numbers of the set M_i . It is clear that M_i consists only of numbers which are multiples of d_i . The number d_i is called the *period of the state* A_i .

Let A_i and A_j be two states belonging to one class. From the foregoing it follows that there exist m and n such that $P_{ij}(m) > 0$ and $P_{ji}(n) > 0$. The number $m+n$ naturally belongs to M_i and, consequently, is divisible by d_i . Let r be an arbitrary and sufficiently large number. Then rd_j belongs to M_j and, hence, $P_{jj}(rd_j) > 0$.

But since

$$P_{ii}(m + rd_j + n) \geq P_{ij}(m) P_{jj}(rd_j) P_{ji}(n)$$

it follows that all numbers of the form $m + rd_j + n$, given sufficiently large r , belong to the set M_i . Since, by the foregoing, the number $m+n$ is divisible by d_i , it follows that rd_j should be divisible by d_i , and since r is arbitrary, d_j should be divisible by d_i . By similar reasoning we can prove that d_i is divisible by d_j . From this it follows that $d_i = d_j$.

Thus, *all states of one and the same class have one and the same period.* (We shall denote it by d .)

The result thus obtained permits us to draw the following conclusion: for two states A_i and A_j belonging to one and the same class, the inequalities $P_{ij}(m) > 0$ and $P_{ji}(n) > 0$ can hold only when m and $-n$ are congruent modulo d .** Thus, if we select a definite state A_α of the class under study, then to each state A_i of this class we can assign a definite number $\beta(i)$ ($\beta(i) = 1, 2, \dots, d$) such that the inequality $P_{\alpha i}(n) > 0$ is possible solely for values of n that satisfy the congruence

* It is easy to notice that M_i contains all the sufficiently large numbers that are multiples of d_i .

** In other words, if the sum $m+n$ is evenly divisible by d .

$n \equiv \beta (i) \pmod{d}$. We combine into a subclass S_β all the states A_i to which the number β has been assigned. To summarize, then, the class of essential states is found to be partitioned into d subclasses S_β . These subclasses possess the property that for each step the system can pass from a state belonging to the subclass S_β into only one of the states of the subclass $S_{\beta+1}$. But if $\beta=d$, then the system passes into one of the states of the subclass S_1 .

Let A_i belong to subclass S_β and A_j to subclass S_γ . From the foregoing it is clear that the probability $P_{ij}(n)$ may be different from zero only when $n \equiv \gamma - \beta \pmod{d}$. But if n satisfies this congruence and is sufficiently great, then the inequality $P_{ij}(n) > 0$ indeed holds.

By way of illustration consider Example 4 of Sec. 17. We see that all the states of the system form one class. Since it is possible, with positive probability, to pass from the state A_i , given any i , to the same state in two steps (and not less than two), it is clear that $d=2$. Thus all the states of the system are subdivided into two subclasses S_1 and S_2 . Put in subclass S_1 all states with odd-numbered subscripts and in subclass S_2 all states with even-numbered subscripts. It is clear that, in one step, it is only possible to pass from each state of the subclass S_1 to a state of the subclass S_2 in the same way as from the subclass S_2 only into the subclass S_1 .

Sec. 19. Theorem on Limiting Probabilities

Theorem. *If for some $s > 0$ all elements of the transition matrix π_s are positive, then there exist constant numbers p_j ($j=1, 2, \dots, k$) such that, irrespective of the subscript i , the equalities*

$$\lim_{n \rightarrow \infty} P_{ij}(n) = p_j$$

hold.

Proof. The idea of the proof of this theorem is exceedingly simple: it is first established that the greatest of the probabilities $P_{ij}(n)$ cannot increase with growth of n and the least cannot decrease. It is then shown that the maximum of the difference $P_{ij}(n) - P_{il}(n)$ ($i, l=1, 2, \dots, k$) tends to zero when $n \rightarrow \infty$. This obviously completes the proof of the theorem. Indeed, by virtue of the well-known theorem on the limit of a monotonic bounded sequence we conclude from the first two indicated properties of the probabilities $P_{ij}(n)$ that there exist

$$\lim_{n \rightarrow \infty} \min_{1 \leq i \leq k} P_{ij}(n) = \bar{p}_j$$

and

$$\lim_{n \rightarrow \infty} \max_{1 \leq i \leq k} P_{ij}(n) = \bar{\bar{p}}_j$$

And since by virtue of the third of the indicated properties

$$\lim_{n \rightarrow \infty} \max_{1 \leq l, l \leq k} |P_{ij}(n) - P_{lj}(n)| = 0$$

it follows that

$$\bar{p}_j = \bar{\bar{p}}_j = p_j$$

Let us now begin to carry out our plan. First of all we notice that for $n > 1$ we have the inequality

$$\begin{aligned} P_{ij}(n) &= \sum_{l=1}^k p_{il} P_{lj}(n-1) \geq \min_{1 \leq l \leq k} P_{lj}(n-1) \sum_{l=1}^k p_{il} = \\ &= \min_{1 \leq l \leq k} P_{lj}(n-1) \quad (1) \end{aligned}$$

This inequality holds for every i , in particular for the one at which

$$P_{ij}(n) = \min_{1 \leq l \leq k} P_{lj}(n)$$

Thus,

$$\min_{1 \leq l \leq k} P_{ij}(n) \geq \min_{1 \leq l \leq k} P_{lj}(n-1)$$

In similar fashion it is easy to notice that

$$\max_{1 \leq l \leq k} P_{ij}(n) \leq \max_{1 \leq l \leq k} P_{lj}(n-1)$$

We can assume that $n > s$, and therefore we have the right to write down, according to (1),

$$P_{ij}(n) = \sum_{r=1}^k P_{ir}(s) \cdot P_{rj}(n-s)$$

We consider the difference

$$\begin{aligned} P_{ij}(n) - P_{lj}(n) &= \sum_{r=1}^k P_{ir}(s) P_{rj}(n-s) - \sum_{r=1}^k P_{lr}(s) P_{rj}(n-s) = \\ &= \sum_{r=1}^k [P_{ir}(s) - P_{lr}(s)] P_{rj}(n-s) \end{aligned}$$

Denote the positive differences $P_{ir}(s) - P_{lr}(s)$ by the symbol $\beta_{il}^{(r)}$, and the nonpositive differences by $\beta'_{il}{}^{(r)}$. Since

$$\sum_{r=1}^k P_{ir}(s) = \sum_{r=1}^k P_{lr}(s) = 1$$

it follows that

$$\sum_{r=1}^k [P_{ir}(s) - P_{lr}(s)] = \sum_{(r)} \beta_{il}^{(r)} - \sum_{(r)} \beta'_{il}{}^{(r)} = 0 \quad (2)$$

From this equation we conclude that

$$h_{il} = \sum_{(r)} \beta_{il}^{(r)} = \sum_{(r)} \beta_{il}'^{(r)}$$

Since by assumption for all i and r ($i, r = 1, 2, 3, \dots, k$) $P_{ir}(s) > 0$, it follows that

$$\sum_{(r)} \beta_{il}'^{(r)} < \sum_{l=1}^k P_{il}(s) = 1$$

And so

$$0 \leq h_{il} < 1$$

Let

$$h = \max_{1 \leq l, l \leq k} h_{il}$$

Since the number of possible outcomes is finite, the quantity h (along with the quantities h_{il}) satisfies the inequalities

$$0 \leq h < 1 \quad (3)$$

From (1) we find that for any i and l ($i, l = 1, 2, \dots, k$)

$$\begin{aligned} |P_{ij}(n) - P_{lj}(n)| &= \left| \sum_{(r)} \beta_{il}^{(r)} P_{rj}(n-s) - \sum_{(r)} \beta_{il}'^{(r)} P_{rj}(n-s) \right| \leq \\ &\leq \left| \max_{1 \leq r \leq k} P_{rj}(n-s) \sum_{(r)} \beta_{il}^{(r)} - \min_{1 \leq r \leq k} P_{rj}(n-s) \sum_{(r)} \beta_{il}'^{(r)} \right| \leq \\ &\leq h \left| \max_{1 \leq r \leq k} P_{rj}(n-s) - \min_{1 \leq r \leq k} P_{rj}(n-s) \right| \leq \\ &\leq h \max_{1 \leq l, l \leq k} |P_{lj}(n-s) - P_{lj}(n-s)| \end{aligned}$$

and consequently, also,

$$\max_{1 \leq l, l \leq k} |P_{ij}(n) - P_{lj}(n)| \leq h \max_{1 \leq l, l \leq k} |P_{lj}(n-s) - P_{lj}(n-s)|$$

Applying this inequality $\left[\frac{n}{s}\right]$ times, we find

$$\begin{aligned} \max_{1 \leq l, l \leq k} |P_{ij}(n) - P_{lj}(n)| &\leq \\ &\leq h^{\left[\frac{n}{s}\right]} \max_{1 \leq l, l \leq k} |P_{lj}\left(n - \left[\frac{n}{s}\right]s\right) - P_{lj}\left(n - \left[\frac{n}{s}\right]s\right)| \end{aligned}$$

Since we always have

$$|P_{ij}(m) - P_{lj}(m)| \leq 1$$

it follows that

$$\max_{1 \leq l, l \leq k} |P_{ij}(n) - P_{lj}(n)| \leq h^{\left[\frac{n}{s}\right]}$$

When $n \rightarrow \infty$ then $\left[\frac{n}{s}\right] \rightarrow \infty$ also; for this reason, by virtue of (3) it follows that

$$\lim_{n \rightarrow \infty} \max_{1 \leq i, l \leq k} |P_{il}(n) - P_{lj}(n)| = 0$$

From what has been proved we also conclude that

$$\sum_{j=1}^k p_j = 1$$

Indeed,

$$\sum_{j=1}^k p_j = \lim_{n \rightarrow \infty} \sum_{j=1}^k P_{ij}(n) = \lim_{n \rightarrow \infty} 1 = 1$$

Thus, we can regard the quantities p_j as probabilities of the occurrence of an outcome $A_j^{(n)}$ in the n th trial when n is great.

The physical meaning of the theorem just proved is clear: the probability of a system being in the state A_j is practically independent of the state that it was in in the remote past.

The above theorem was first proved by the creator of the theory of chain dependences A. A. Markov. It was the first rigorously proved result of the so-called ergodic theorems that play an important role in modern physics.

It may be proved that if the possible states of a system form a single essential class, then the ergodic theorem holds.

Sec. 20. Generalizing the DeMoivre-Laplace Theorem to a Sequence of Chain-Dependent Trials

We shall now focus our attention on a sequence of trials, in each of which an event E may or may not occur. We shall assume that the trials are not independent, but are connected into a simple Markov chain. Thus, if in the k th trial the event E occurred, then the probability that in the next $(k+1)$ st trial event E will again occur is α ; now the probability that event E will occur in the $(k+1)$ st trial, given that in the k th trial event \bar{E} occurred, is β . Hence, in our case, the transition probabilities are given by the matrix

$$\begin{pmatrix} \alpha & 1 - \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

We shall henceforward assume that both α and β are different from 0 and 1, for these cases are of no particular interest. The scheme at hand is understandably a natural generalization of the scheme of independent trials proposed by James Bernoulli and examined by us in the preceding chapter.

We must note that assigning the transition matrix does not completely specify the system of trials, because the first trial has no precedent and, consequently, the probabilities of occurrence of events E and \bar{E} in the first trial are unknown to us. We therefore denote by p_1 the probability of occurrence of E in the first trial and by $q_1=1-p_1$ the probability of event \bar{E} occurring in the first trial.

We first solve the two following problems: (1) to find the probability that event E will occur in the k th trial; (2) to find the probability that E will occur in the j th trial if event E occurred in the i th trial ($i < j$).

Denote by p_k the probability that event E will occur in the k th trial and put $q_k=1-p_k$. It is obvious that in the k th trial E can occur in two mutually exclusive ways: event E will occur in the $(k-1)$ st trial and will occur again in the next trial; event \bar{E} will occur in the $(k-1)$ st trial and event E will occur in the next trial. Using the formula for total probability, we find that

$$p_k = p_{k-1}\alpha + q_{k-1}\beta$$

Since $q_{k-1}=1-p_{k-1}$, then setting $\delta=\alpha-\beta$ we find

$$p_k = p_{k-1}\delta + \beta$$

In particular, when $k=2$

$$p_2 = p_1\delta + \beta$$

When $k=3$

$$p_3 = p_2\delta + \beta = p_1\delta^2 + \beta(1+\delta)$$

It is easy to verify that for any $k > 1$

$$p_k = p_1\delta^{k-1} + \beta(1 + \delta + \dots + \delta^{k-2}) = \left(p_1 - \frac{\beta}{1-\delta}\right)\delta^{k-1} + \frac{\beta}{1-\delta} \quad (1)$$

Given the assumptions we have made relative to α and β , the quantity δ satisfies the inequality $|\delta| < 1$. From the preceding formula it follows that as $k \rightarrow \infty$

$$p_k \rightarrow \frac{\beta}{1-\delta}$$

It is interesting to note that the constant to which p_k tend does not depend on the probability p_1 .

Since the quantity $\frac{\beta}{1-\delta}$ plays the role of a "limiting probability", it is natural to introduce the notation

$$p = \frac{\beta}{1-\delta} = \frac{\beta}{1-\alpha+\beta}, \quad q = 1-p = \frac{1-\alpha}{1-\delta}$$

In this notation,

$$p_k = p + (p_1 - p)\delta^{k-1} \quad (1')$$

Now denote by $p_j^{(i)}$ the probability of event E occurring in the j th trial if it occurred in the i th trial. Proceeding as we have just done, it will become clear that the probabilities $p_j^{(i)}$ satisfy the difference equation

$$p_j^{(i)} = p_{j-1}^{(i)}\delta + \beta$$

for all $j > i + 1$. But $p_{i+1}^{(i)} = \alpha$, and therefore, applying the procedure just utilized, we find

$$p_j^{(i)} = \alpha\delta^{j-i-1} + \beta(1 + \delta + \dots + \delta^{j-i-2}) = \frac{\beta}{1-\delta} + \frac{1-\alpha}{1-\delta}\delta^{j-i} \quad (2)$$

or

$$p_j^{(i)} = p + q\delta^{j-i} \quad (2')$$

Let us now look for the probability of the m -fold occurrence of event E among n trials. For this purpose, we break up the desired probability, which we continue to denote by $P_n(m)$, into four summands:

$$P_n(m) = P_n(m, EE) + P_n(m, E\bar{E}) + P_n(m, \bar{E}E) + P_n(m, \bar{E}\bar{E})$$

The first term signifies the probability of the m -fold occurrence of event E in n trials on condition that in the first trial and last trial the event E will occur. The meaning of the other notations is now clear without any further explanations. To evaluate $P_n(m, EE)$ we first consider the following arrangement of trial results:

in the first r_1 trials, events E occurred

then in s_1 trials, events \bar{E} occurred

"	"	r_2	"	"	E	"
.
"	"	s_{k-1}	"	"	\bar{E}	"
"	"	r_k	"	"	E	"

As will readily be seen, the probability of such an outcome is

$$p_1\alpha^{r_1-1}(1-\alpha)(1-\beta)^{s_1-1}\beta\dots\beta\alpha^{r_k-1} =$$

$$= p_1\alpha^{r_1+\dots+r_k-k}(1-\alpha)^{k-1}(1-\beta)^{s_1+\dots+s_{k-1}-k+1}\beta^{k-1}$$

But since

$$\sum_{i=1}^k r_i = m, \quad \sum_{i=1}^{k-1} s_i = n - m$$

this probability is

$$p_1\alpha^{m-k}(1-\alpha)^{k-1}(1-\beta)^{n-m-k+1}\beta^{k-1}$$

Note that it is dependent solely on m , n and k and is independent of the values of r_j and s_j . Since the number m may be decomposed

into k positive summands in C_{m-1}^{k-1} ways, and $n-m$ may be represented in the form of a $k-1$ positive summand in C_{n-m-1}^{k-2} ways, the probability of the m -fold occurrence of event E , in which events E will occur in the form of k groups and events \bar{E} in the form of $k-1$ groups, is

$$C_{m-1}^{k-1} C_{n-m-1}^{k-2} p_1 \alpha^{m-k} (1-\alpha)^{k-1} (1-\beta)^{n-m-k+1} \beta^{k-1}$$

Since k can take on any value from 2 to m ,

$$P_n(m, EE) = p_1 \sum_{k=2}^m C_{m-1}^{k-1} \alpha^{m-k} (1-\alpha)^{k-1} C_{n-m-1}^{k-2} (1-\beta)^{n-m-k+1} \beta^{k-1}$$

In similar fashion we find:

$$P_n(m, E\bar{E}) = p_1 \sum_{k=1}^m C_{m-1}^{k-1} \alpha^{m-k} (1-\alpha)^k C_{n-m-1}^{k-1} (1-\beta)^{n-m-k} \beta^{k-1}$$

$$P_n(m, \bar{E}E) = q_1 \sum_{k=1}^m C_{m-1}^{k-1} \alpha^{m-k} (1-\alpha)^{k-1} C_{n-m-1}^{k-1} (1-\beta)^{n-m-k} \beta^k$$

$$P_n(m, \bar{E}\bar{E}) = q_1 \sum_{k=2}^m C_{m-1}^{k-2} \alpha^{m-k+1} (1-\alpha)^{k-1} C_{n-m-1}^{k-1} (1-\beta)^{n-m-k} \beta^k$$

To evaluate all these four probabilities, consider the expression

$$A_{mn} = \sum_{k=1}^m C_m^k \alpha^{m-k} (1-\alpha)^k C_{n-m}^k (1-\beta)^{n-m-k} \beta^k$$

and introduce the notations

$$m = np + z \sqrt{\frac{m\alpha(1-\alpha) + (n-m)\beta(1-\beta)}{(1-\alpha+\beta)^2}} \quad (3)$$

and

$$k = m(1-\alpha) + u \sqrt{m\alpha(1-\alpha)}$$

$$k = (n-m)\beta + v \sqrt{(n-m)\beta(1-\beta)}$$

We will carry out the computations assuming that

$$u = o(m^{1/6}), \quad v = o(n^{1/6})$$

where γ is some number that satisfies the inequalities $0 < \gamma < 1/6$.

We decompose the quantity A_{mn} into three summands:

$$A_{mn} = \Sigma_1 + \Sigma_2 + \Sigma_3$$

putting

$$\begin{aligned}\Sigma_1 &= \sum_{k=1}^{m(1-\alpha)-u_1} \frac{V \overline{m\alpha(1-\alpha)}}{V \overline{m\alpha(1-\alpha)}}, & \Sigma_2 &= \sum_{m(1-\alpha)-u_1}^{m(1-\alpha)+u_1} \frac{V \overline{m\alpha(1-\alpha)}}{V \overline{m\alpha(1-\alpha)}} \\ \Sigma_3 &= \sum_{m(1-\alpha)+u_1}^m \frac{V \overline{m\alpha(1-\alpha)}}{V \overline{m\alpha(1-\alpha)}}\end{aligned}$$

We begin the computations with the middle sum.

Repeating word-for-word the arguments given in the proof of the local theorem of DeMoivre-Laplace, we find

$$\begin{aligned}C_m^k \alpha^{m-k} (1-\alpha)^k &= \frac{1}{\sqrt{2\pi\alpha(1-\alpha)m}} e^{-\frac{u^2}{2}} (1 + \omega'_n) \\ C_{n-m}^k \beta^k (1-\beta)^{n-m-k} &= \frac{1}{\sqrt{2\pi(n-m)\beta(1-\beta)}} e^{-\frac{v^2}{2}} (1 + \omega''_n)\end{aligned}$$

The quantities ω'_n and ω''_n approach zero uniformly within the chosen bounds.

Thus,

$$\Sigma_2 = \frac{1}{2\pi \sqrt{m(n-m)\alpha\beta(1-\alpha)(1-\beta)}} \sum_{u=-u_1}^{u_1} e^{-\frac{u^2+v^2}{2}} (1 + \omega'_n) (1 + \omega''_n)$$

According to the DeMoivre-Laplace integral theorem,

$$\Sigma_2 = \frac{1}{2\pi \sqrt{(n-m)\beta(1-\beta)}} \int_{-u_1}^{u_1} e^{-\frac{u^2+v^2}{2}} du (1 + \omega_n)$$

Since u_1 tends to ∞ together with n , we can write

$$\Sigma_2 = \frac{1}{2\pi \sqrt{(n-m)\beta(1-\beta)}} \int_{-\infty}^{\infty} e^{-\frac{u^2+v^2}{2}} du (1 + \bar{\omega}_n)$$

But u and v are connected by the equation

$$m(1-\alpha) + u \sqrt{m\alpha(1-\alpha)} = (n-m)\beta + v \sqrt{(n-m)\beta(1-\beta)}$$

Substitute here the value of m from (3). After obvious simplifications we find

$$\begin{aligned}z \sqrt{m\alpha(1-\alpha) + (n-m)\beta(1-\beta)} + u \sqrt{m\alpha(1-\alpha)} &= \\ &= v \sqrt{(n-m)\beta(1-\beta)}\end{aligned}$$

Whence

$$v = \frac{1}{\sqrt{(n-m)\beta(1-\beta)}} \left[z \sqrt{m(1-\alpha)\alpha + (n-m)\beta(1-\beta)} + u \sqrt{m\alpha(1-\alpha)} \right]$$

Thus,

$$u^2 + v^2 = z^2 + \frac{m\alpha(1-\alpha) + (n-m)\beta(1-\beta)}{(n-m)\beta(1-\beta)} \left[u + z \sqrt{\frac{m\alpha(1-\alpha)}{m\alpha(1-\alpha) + (n-m)\beta(1-\beta)}} \right]^2$$

and, consequently,

$$\sum_2 = \frac{1}{\sqrt{2\pi [m\alpha(1-\alpha) + (n-m)\beta(1-\beta)]}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}_n)$$

We now note that according to (3)

$$m\alpha(1-\alpha) + (n-m)\beta(1-\beta) = np\alpha(1-\alpha) + nq\beta(1-\beta) + O(z\sqrt{n}) = npq(1+\alpha-\beta)(1-\alpha+\beta) + O(z\sqrt{n})$$

Thus, asymptotically,

$$m\alpha(1-\alpha) + (n-m)\beta(1-\beta) = npq(1+\alpha-\beta)(1-\alpha+\beta)$$

and

$$\sum_2 = \frac{1}{\sqrt{2\pi npq(1+\alpha-\beta)(1-\alpha+\beta)}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}'_n)$$

To estimate the sum \sum_1 , we introduce the notation

$$u_i = C_m^i \alpha^{m-i} (1-\alpha)^i C_{n-m}^i (1-\beta)^{n-m-i} \beta^i$$

and note that the relation

$$\frac{u_i}{u_{i+1}} = \frac{(i+1)^2}{(m-i)(n-m-i)} \frac{(1-\beta)\alpha}{(1-\alpha)\beta}$$

increases with increasing i and remains less than unity for i that are not too large. Let $j = m(1-\alpha) - u_1 \sqrt{m\alpha(1-\alpha)}$,

$$v_j = u_j, \quad v_{j-1} = u_{j-1}, \quad \frac{v_{j-1}}{v_j} = \kappa$$

and for the remaining values of i ,

$$v_i = v_j \kappa^{j-i}$$

It is clear that

$$\sum_1 < v_1 + v_2 + \dots + v_j < v_j \frac{1}{1-\kappa}$$

Since, in accordance with calculations carried out earlier,

$$v_j = \frac{1}{2\pi \sqrt{m(n-m)\alpha\beta(1-\alpha)(1-\beta)}} e^{-\frac{u_1^2 + v_1^2}{2}} (1 + \omega'_n)(1 + \omega''_n)$$

and for sufficiently large n

$$\kappa < \frac{1}{2}$$

it follows that

$$\sum_i = o(1)$$

In similar fashion it becomes evident that

$$\sum_s = o(1)$$

As a result we find that \sum_2 is the main part of A_{mn} .

Comparing A_{mn} with the sought-for probabilities, we find:

$$P_n(m, EE) = \frac{p_1\beta}{\sqrt{2\pi[m\alpha(1-\alpha) + (n-m)\beta(1-\beta)]}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}_n)$$

$$P_n(m, E\bar{E}) = \frac{p_1(1-\alpha)}{\sqrt{2\pi[m\alpha(1-\alpha) + (n-m)\beta(1-\beta)]}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}_n)$$

$$P_n(m, \bar{E}E) = \frac{q_1\beta}{\sqrt{2\pi[m\alpha(1-\alpha) + (n-m)\beta(1-\beta)]}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}_n)$$

$$P_n(m, \bar{E}\bar{E}) = \frac{q_1(1-\alpha)}{\sqrt{2\pi[m\alpha(1-\alpha) + (n-m)\beta(1-\beta)]}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}_n)$$

From this we conclude that

$$P_n(m) = \frac{1}{\sqrt{2\pi npq \frac{1+\alpha-\beta}{1-\alpha+\beta}}} e^{-\frac{z^2}{2}} (1 + \bar{\omega}'_n)$$

The local theorem is proved.

We note that if the transition probabilities satisfy the equality

$$\alpha = \beta$$

then the local theorem assumes the same form as in the case of independent trials.

Proceeding in the usual way we can also derive the integral limit theorem from the local theorem, no matter what z_1 and z_2 are,

$$P \left\{ z_1 \leq \frac{m-np}{\sqrt{npq \frac{1+\alpha-\beta}{1-\alpha+\beta}}} < z_2 \right\} = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{z^2}{2}} dz + \omega_n$$

The quantity ω_n tends to zero uniformly in z_1 and z_2 as n increases to infinity.

EXERCISES

1. The transition probabilities are given by the matrix

$$\pi_1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

What is the number of the states? Find the two-step transition probabilities from state to state.

2. An electron may reside in one of a countable set of orbits depending on its energy. Transition from the i th orbit to the j th orbit takes place in one second with a probability $c_i e^{-\alpha|i-j|}$. Find: (a) the transition probabilities for two seconds; (b) the constants c_i .

3. The transition probabilities are given by the matrix

$$\pi_1 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Is Markov's ergodic theorem applicable in this case? If it is, then find the limiting probabilities.

CHAPTER 4

Random Variables and Distribution Functions

Sec. 21. Basic Properties of Distribution Functions

One of the basic concepts of probability theory is that of the random variable. Before giving a formal definition we shall illustrate it with a number of examples.

The number of cosmic particles impinging on some area of the earth's surface during a definite time interval is subject to appreciable variations depending on a multitude of random factors.

The number of calls arriving from subscribers at a telephone exchange during a definite time interval is also a random variable and takes on values of one kind or another depending on accidental circumstances.

The deviation of the point of impact of a shell from the centre of a target is determined by a large number of diversified causes of an accidental nature. The result is that in the theory of gunfire one has to consider the dispersion of shells about the centre of the target as a random phenomenon and regard the indicated deviations as random variables.

The velocity of a gas molecule does not remain invariable but changes depending on collisions with other molecules, of which there are great numbers even within a very brief span of time. Knowing the molecule velocity at a given instant, one cannot state with full definiteness what its value will be, say 0.01 or 0.001 second hence. Change of molecular velocity is of a random nature.

These examples show very definitely that random variables are involved in the most diversified fields of science and technology. The natural and extremely important problem arises of creating methods for studying random variables.

Despite the diversity of concrete content in the foregoing examples, they essentially present the same picture from the viewpoint of the mathematician. Namely, in each instance we have to do with a quantity that in one way or another describes the phenomenon under study.

Under the effect of random circumstances, each of these quantities is capable of taking on a variety of values. One cannot state beforehand what value the quantity will assume, for it varies in random fashion from trial to trial.

Therefore, in order to know a random variable it is first and foremost necessary to know the values that it can assume. However, a simple list of values of the random variable is not enough for us to draw essential conclusions. Indeed, if in the third example we consider a gas at different temperatures, the possible values of molecular velocities will remain the same, whereas the states of the gas will differ. Thus, to specify a random variable it is necessary to know not only what values it can assume, but also how often, that is, with what probability, it assumes these values.

The diversity of random variables is extremely great. The number of assumed values may be finite, countable and uncountable; the values may be distributed discretely or fill the intervals continuously, or not fill the intervals, but be spread out, everywhere dense. In order to specify the probabilities of the values of random variables that are so diversified, and to be able to specify them in one and the same fashion, we introduce into the theory of probability the concept of the *distribution function of a random variable*.

Let ξ be a random variable and x be an arbitrary real number. The probability that ξ will take on a value less than x is called the *distribution function of probabilities* of the random variable ξ :

$$F(x) = P\{\xi < x\}$$

Let us agree from now on, as a rule, to denote random variables by *Greek* letters and the values that they assume by *lower-case Latin* letters.

Let us summarize what has been said, remaining at the level of a qualitative description: a *random variable* is a variable quantity whose values depend on chance and for which a distribution function of probabilities has been defined.*

We consider examples of distribution functions.

Example 1. Denote by μ the number of occurrences of event A in a sequence of n independent trials, in each of which the probability of its occurrence is a constant equal to p . Depending on chance, μ can assume all integral values from 0 to n inclusive. According to the results of Chapter 2,

$$P_n(\mu) = P\{\mu = m\} = C_n^m p^m q^{n-m}$$

* A formal definition of a random variable will be given on p. 127.

The distribution function of the variable μ is defined in the following manner:

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sum_{k \leq x} P_n(k) & \text{for } 0 < x \leq n \\ 1 & \text{for } x > n \end{cases}$$

The distribution function is a step-like line with jumps at the points $x=0, 1, 2, \dots, n$; the jump at the point $x=k$ is $P_n(k)$.

The foregoing example shows that the so-called Bernoulli scheme may be included in the general theory of random variables.

Example 2. Let the random variable ξ take on the values $0, 1, 2, \dots$ with the probabilities

$$p_n = \mathbf{P} \{ \xi = n \} = \frac{\lambda^n e^{-\lambda}}{n!} \quad (n = 0, 1, 2, \dots)$$

where $\lambda > 0$ is a constant. The distribution function of ξ is in the form of a sort of staircase with an infinite number of steps, with jumps at all nonnegative integral points. The magnitude of the jump at the point $x=n$ is equal to p_n ; for $x \leq 0$ we have $F(x)=0$. Of the random variable examined in this case, it is said that it is distributed in accordance with the *Poisson law*.

Example 3. We say that a random variable is *normally distributed* if its distribution function has the form

$$\Phi(x) = C \int_{-\infty}^x e^{-\frac{(z-a)^2}{2\sigma^2}} dz$$

where $C > 0$, $\sigma > 0$, and a are constants. Later, we will establish the relationship between the constants σ and C and will elucidate the probabilistic meaning of the parameters a and σ . Normally distributed random variables play a particularly important role in probability theory and its applications; we will have good reason to be convinced of this later on.

Note that if in the first two examples considered above the random variable could take on only a finite or countable set of values (*discrete variables*), random variables distributed in accord with the normal law can assume values from any interval. Indeed, as we shall see below, the probability of a normally distributed random variable taking on a value lying in the interval $x_1 \leq \xi < x_2$ is equal to

$$\Phi(x_2) - \Phi(x_1) = C \int_{x_1}^{x_2} e^{-\frac{(z-a)^2}{2\sigma^2}} dz$$

and consequently is positive for any x_1 and x_2 ($x_1 \neq x_2$).

Now, after these preliminary remarks of an intuitive nature, we pass to a rigorous formal exposition of the notion of a random variable.

In defining a random variable we shall proceed in accordance with the general concept of a random event of a set of elementary events U , a set of random events and a probability measure $\mathbf{P}\{A\}$ defined on it. In other words, our point of departure is a probability space $\{U, F, \mathbf{P}\}$. With each elementary event e we associate a certain number

$$\xi = f(e)$$

We say that ξ is a random variable if the function $f(e)$ is measurable relative to the probability introduced into the set U under consideration. To put it otherwise, we demand that for each value of x ($-\infty < x < +\infty$) the set A_x of those e for which $f(e) < x$ should belong to the set F of random events and, hence, that for it there should be defined the probability

$$\mathbf{P}\{\xi < x\} = \mathbf{P}\{A_x\} = F(x)$$

which we have called the *distribution function* of the random variable ξ .

Example 4. We consider a sequence n of independent trials in each of which the probability of occurrence of the event A is constant and equal to p . In this example, the elementary events consist of sequences of occurrences and nonoccurrences of the event A in n trials. Thus, one of the elementary events will be the occurrence of event A in each of the trials. It is easy to compute that there will be a total of 2^n elementary events.

We define the function $\mu = f(e)$ of an elementary event e as follows: it is equal to the number of occurrences of the event A in the elementary event e . According to the results of Chapter 2,

$$\mathbf{P}\{\mu = k\} = P_n(k) = C_n^k p^k q^{n-k}$$

The measurability of the function $\mu = f(e)$ in the probability field is immediately obvious, whence, by definition, we conclude that μ is a random variable.

Example 5. Three observations are taken of the position of a molecule moving in a straight line. The set of elementary events consists of the points of three-dimensional Euclidean space R_3 . The set of random events F consists of all possible Borel sets in the space R_3 .

For each random event A the probability $\mathbf{P}(A)$ is defined by the equation

$$\mathbf{P}(A) = \frac{1}{(\sigma \sqrt{2\pi})^3} \int \int \int e^{-\frac{1}{2\sigma^2} [(x_1 - a)^2 + (x_2 - a)^2 + (x_3 - a)^2]} dx_1 dx_2 dx_3$$

Now consider the function $\xi = f(e)$ of the elementary event $e = (x_1, x_2, x_3)$ defined by the equation

$$\xi = \frac{1}{3} (x_1 + x_2 + x_3)$$

This function is measurable relative to the probability we introduced, and so ξ is a random variable. Its distribution function is

$$\begin{aligned} F(x) = P\{\xi \leq x\} &= \frac{1}{(\sigma \sqrt{2\pi})^3} \int \int \int_{x_1 + x_2 + x_3 \leq 3x} e^{-\frac{1}{2\sigma^2} \sum_{k=1}^3 (x_k - a)^2} dx_1 dx_2 dx_3 = \\ &= \frac{1}{\sigma \sqrt{\frac{2}{3}\pi}} \int_{-\infty}^x e^{-\frac{3(z-a)^2}{2\sigma^2}} dz \end{aligned}$$

From the point of view just developed, operations on random variables reduce to familiar operations on functions. Thus, if ξ_1 and ξ_2 are random variables, that is, measurable functions relative to the probability introduced,

$$\xi_1 = f_1(e), \quad \xi_2 = f_2(e)$$

then any Borel function of these variables is also a random variable. To illustrate,

$$\zeta = \xi_1 + \xi_2$$

is measurable relative to the introduced probability and for this reason is a random variable.

In Sec. 24 we will develop this remark and will derive a number of results that are important in theory and applications. In particular, a formula will be derived for the distribution function of a sum based on the distribution of the summands.

With the aid of the distribution function of the random variable ξ it is possible to define the probability of the inequality $x_1 \leq \xi < x_2$ for any x_1 and x_2 . Indeed, if by A we denote an event that consists in ξ taking on a value less than x_2 , by B an event consisting in the fact that $\xi < x_1$ and, finally, by C the event that $x_1 \leq \xi < x_2$, then it is obvious that the following equation holds:

$$A = B + C$$

Since events B and C are mutually exclusive, it follows that

$$P(A) = P(B) + P(C)$$

But

$$P(A) = F(x_2), \quad P(B) = F(x_1), \quad P(C) = P\{x_1 \leq \xi < x_2\}$$

therefore

$$\mathbf{P}\{x_1 \leq \xi < x_2\} = F(x_2) - F(x_1) \quad (1)$$

Since, by definition, probability is a nonnegative number, it follows from (1) that for any x_1 and x_2 ($x_2 > x_1$) we have the inequality

$$F(x_1) \leq F(x_2)$$

that is, the *distribution function of a random variable is a nondecreasing function*.

It is further obvious that the distribution function $F(x)$ for any x satisfies the inequality

$$0 \leq F(x) \leq 1 \quad (2)$$

We say that at $x=x_0$ the distribution function $F(x)$ has a *jump* if

$$F(x_0+0) - F(x_0-0) = C_0 > 0$$

A distribution function cannot have more than a countable set of jumps. Indeed, a distribution function cannot have more than one jump of magnitude greater than $1/2$, more than three of magnitude from one fourth to one half ($1/4 < C_0 \leq 1/2$). Generally, there can be no more than $2^n - 1$ jumps of magnitude from $\frac{1}{2^n}$ to $\frac{1}{2^{n-1}}$. It is quite clear that we can number all the jumps, arranging them in magnitude beginning with large values and repeating equal values as many times as the function $F(x)$ has jumps of that magnitude.

We shall now establish some other general properties of distribution functions. We define $F(-\infty)$ and $F(+\infty)$ by the equations

$$F(-\infty) = \lim_{n \rightarrow +\infty} F(-n), \quad F(+\infty) = \lim_{n \rightarrow \infty} F(+n)$$

and will prove that

$$F(-\infty) = 0, \quad F(+\infty) = 1$$

Indeed, since the inequality $\xi < +\infty$ is certain, it follows that

$$\mathbf{P}\{\xi < +\infty\} = 1$$

Denote by Q_k the event that $k-1 \leq \xi < k$. Since the event $\xi < +\infty$ is equivalent to the sum of events Q_k , it follows, on the basis of the extended addition axiom, that

$$\mathbf{P}\{\xi < +\infty\} = \sum_{k=-\infty}^{\infty} \mathbf{P}\{Q_k\}$$

Consequently, as $n \rightarrow \infty$,

$$\sum_{k=1-n}^n \mathbf{P}\{Q_k\} = \sum_{k=1-n}^n [F(k) - F(k-1)] = F(n) - F(-n) \rightarrow 1$$

From this, taking into account the inequality (2), we conclude that as $n \rightarrow \infty$

$$F(-n) \rightarrow 0, F(+n) \rightarrow 1$$

The distribution function is continuous on the left.

Choose some increasing sequence $x_0 < x_1 < x_2 < \dots < x_n < \dots$ converging to x .

Denote by A_n the event $\{x_n \leq \xi < x\}$. It is then clear that $A_i \subset A_j$, for $i > j$, and the product of all the events A_n is an impossible event. On the basis of the continuity axiom, we should have

$$\begin{aligned} \lim_{n \rightarrow \infty} P(A_n) &= \lim_{n \rightarrow \infty} \{F(x) - F(x_n)\} = F(x) - \lim_{n \rightarrow \infty} F(x_n) = \\ &= F(x) - F(x-0) = 0 \end{aligned}$$

which is what we set out to prove.

In exactly the same way it can be proved that

$$P\{\xi \leq x\} = F(x+0)$$

We see, therefore, that *every distribution function is a nondecreasing function that is continuous on the left and satisfies the conditions $F(-\infty)=0$ and $F(+\infty)=1$* . The converse is also true: *every function that satisfies the enumerated conditions may be regarded as the distribution function of some random variable*.

We note that whereas every random variable uniquely defines its distribution function, there are an arbitrary number of different random variables having one and the same distribution function. Thus, if ξ takes on two values -1 and $+1$, each with a probability $1/2$ and $\eta = -\xi$, then it is clear that ξ is always different from η . Nevertheless, both these random variables have one and the same distribution function

$$F(x) = \begin{cases} 0 & \text{for } x \leq -1 \\ \frac{1}{2} & \text{for } -1 < x \leq 1 \\ 1 & \text{for } x > 1 \end{cases}$$

Sec. 22. Continuous and Discrete Distributions

The behaviour of a random variable is sometimes described not by specifying its distribution function but in some other way. Any such description is called a *distribution law* of the random variable if by following specific rules it is possible to obtain the distribution function from it. For instance, the interval function $P\{x_1, x_2\}$, which is the probability of the inequality $x_1 \leq \xi < x_2$, is such a distribution law. Indeed, knowing $P\{x_1, x_2\}$ we can find the distribution function from the formula

$$F(x) = P\{-\infty, x\}$$

We already know that it is also possible from $F(x)$ to find the function $P\{x_1, x_2\}$ for any x_1 and x_2 :

$$P\{x_1, x_2\} = F(x_2) - F(x_1)$$

As a distribution law, it is often useful to take the set function $P\{E\}$ defined for all Borel sets and representing the probability that the random variable ξ will take on a value belonging to the set E . By virtue of the extended addition axiom, the probability $P\{E\}$ is a completely additive set function, that is, for any set E , which is the union of a finite or countable number of disjoint sets E_k ,

$$P\{E\} = \sum P\{E_k\}$$

Of all possible random variables we isolate first of all those which can assume only a finite or countable set of values. We call these variables *discrete*. For a complete probabilistic description of a discrete random variable which with positive probabilities takes on the values x_1, x_2, x_3, \dots , it is sufficient to know the probabilities $p_k = P\{\xi = x_k\}^*$. It is obvious that by using the probabilities p_k it is possible to define the distribution function $F(x)$ by means of the equation

$$F(x) = \sum p_k$$

in which the summation is extended over all indices for which $x_k < x$.

The distribution function of any discrete variable is discontinuous and increases in jumps for those values of x which are possible values of ξ . The magnitude of the jumps of the function $F(x)$ at the point x is, as we found out earlier, equal to the difference $F(x+0) - F(x)$.

If two possible values of the variable ξ are separated by an interval in which there are no other possible values of ξ , then the distribution function $F(x)$ is constant in this interval. If the possible values of ξ is a finite number, say n , then the distribution function $F(x)$ is a step-like curve with an $n+1$ interval of constancy. But if there is a countable set of possible values of ξ , then this set may also be everywhere dense so that there may not be any intervals of constancy in the distribution function of the discrete random variable. By way of illustration, let the possible values of ξ be all of the rational numbers and only them. Let all these numbers be numbered in some fashion: r_1, r_2, \dots , and let the probabilities $P\{\xi = r_k\} = p_k$ be defined by means of the equation $p_k = \frac{1}{2^k}$. In our example, all rational points are points of discontinuity of the distribution function.

As another important class of random variables we isolate those for which there is a nonnegative function $p(x)$ that satisfies the fol-

* These, and only these, values x_n will be called *possible values* of the discrete random variable ξ .

lowing equation for any x :

$$F(x) = \int_{-\infty}^x p(z) dz$$

Random variables that possess this property are called *continuous*; the function $p(x)$ is called the *density of probability distribution* or the *probability density function*.

If the function $F(x)$ is absolutely continuous, and all the more so if it is differentiable for all values of the argument, then its derivative is the density function: $p(x) = F'(x)$.

We note that the density function has the following properties:

- (1) $p(x) \geq 0$;
- (2) for any x_1 and x_2 it satisfies the equation

$$\mathbf{P} \{x_1 \leq \xi < x_2\} = \int_{x_1}^{x_2} p(x) dx$$

in particular, if $p(x)$ is continuous at the point x , then to within higher-order infinitesimals $\mathbf{P} \{x \leq \xi < x + dx\} = p(x) dx$;

$$(3) \int p(x) dx = 1.$$

Quantities distributed in accord with the normal or the uniform law* are instances of continuous random variables.

Example. Let us examine the normal distribution law more closely. For it the density function is

$$p(x) = C \cdot e^{-\frac{(x-a)^2}{2\sigma^2}}$$

The constant C is determined on the basis of Property 3. Indeed,

$$C \int e^{-\frac{(x-a)^2}{2\sigma^2}} dx = 1$$

Changing the variables $\frac{x-a}{\sigma} = z$ reduces this equation to the form

$$C\sigma \int e^{-\frac{z^2}{2}} dz = 1$$

The integral on the left-hand side of this equation is known as the *Poisson integral*; here,

$$\int e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

* This is a law with a distribution function varying lineary from 0 to 1 in some interval (a, b) and equal to zero left of the point a and equal to one to the right of b .

We thus find that

$$C = \frac{1}{\sigma \sqrt{2\pi}}$$

and, consequently, for the normal distribution

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

The function $p(x)$ reaches a maximum at $x=a$ and has points of inflexion at $x=a\pm\sigma$; the axis of abscissas serves it as an asymptote as $x\rightarrow\pm\infty$. To illustrate the effect of the parameter σ on the shape

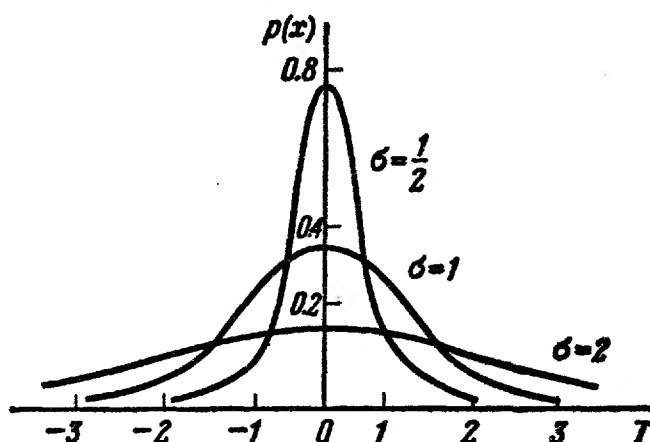


Fig. 14

of the graph of the normal density function, we give (in Fig. 14) the graphs of $p(x)$ for $a=0$ and (1) $\sigma^2 = \frac{1}{4}$; (2) $\sigma^2 = 1$; (3) $\sigma^2 = 4$. We see that the smaller the value of σ , the greater the maximum value of the curve $p(x)$ and the steeper the drop. For one thing, this means that the probability of falling in the interval $(-\alpha, \alpha)$ is greater for the normally distributed random variable (with parameter $a=0$) for which the quantity σ is smaller. Hence, we can consider σ a characteristic of the dispersion of the values of the variable ξ . For $a \neq 0$, the density curves have the same shape but are shifted to the right ($a > 0$) or to the left ($a < 0$) depending on the sign of the parameter a .

Of course, there are random variables other than discrete and continuous. Besides those that behave in one set of intervals like continuous variables and in others like discrete variables, there are variables which are neither discrete nor continuous in any interval. In this group of random variables are all those functions whose distributions are continuous but which only increase on a set of Lebesgue measure zero. An example of such a random variable is the quantity having the well-known Cantor curve as its distribution function. Let us recall the construction of this curve. The variable ξ only takes on values between zero and unity. Thus its distribution function

satisfies the equations

$$F(x)=0 \text{ for } x \leq 0, \quad F(x)=1 \text{ for } x > 1$$

Within the interval $(0, 1)$, ξ assumes the values only in the first and the last third, in each with a probability $1/2$. Thus,

$$F(x) = \frac{1}{2} \text{ for } \frac{1}{3} < x \leq \frac{2}{3}$$

In the intervals $(0, \frac{1}{3})$ and $(\frac{2}{3}, 1)$, ξ again can assume values only in the first and the last third of each of them, and in each with a probability $1/4$. This defines the values of $F(x)$ in two more intervals:

$$F(x) = \frac{1}{4} \text{ for } \frac{1}{9} < x \leq \frac{2}{9}$$

$$F(x) = \frac{3}{4} \text{ for } \frac{7}{9} < x \leq \frac{8}{9}$$

In each of the remaining intervals the same construction is repeated; this process continues ad infinitum. As a result, the function $F(x)$ proves to be defined on a countable set of intervals which are cointervals of a certain nowhere-dense perfect set of measure zero. On this set we redefine the function $F(x)$ relative to continuity. The variable ξ with a thus defined distribution function is not discrete, for its distribution function is continuous, but at the same time ξ is not continuous, for its distribution function is not the integral of its derivative.

All the definitions that we have introduced are readily carried over to the case of conditional probabilities. Thus, for instance, if the event B is such that $\mathbf{P}\{B\} > 0$, then we shall call the function $F(x/B) = \mathbf{P}\{\xi < x/B\}$ the *conditional distribution function* of the random variable ξ , given the condition B . It is obvious that $F(x/B)$ possesses all the properties of an ordinary distribution function.

Sec. 23. Multidimensional Distribution Functions

In what follows we will need, in addition to the notion of a random variable, also the concept of a random vector or, as it is often called, a multidimensional random variable.

Let us consider a probability space $\{U, F, \mathbf{P}\}$ on which are defined n random variables:

$$\xi_1 = f_1(e), \quad \xi_2 = f_2(e), \quad \dots, \quad \xi_n = f_n(e)$$

The vector $(\xi_1, \xi_2, \dots, \xi_n)$ is called an *n -dimensional random variable*.

Let $(\xi_1, \xi_2, \dots, \xi_n)$ be a random vector. Denote by $\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n\}$ the set of elementary events e for which all the following

inequalities hold at the same time: $f_1(e) < x_1, f_2(e) < x_2, \dots, f_n(e) < x_n$. Inasmuch as this event is the product of events $\{f_1(e) < x_1\}, \{f_2(e) < x_2\}, \dots, \{f_n(e) < x_n\}$, it belongs to the set F , that is,

$$\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n\} \in F$$

Thus, for any set of numbers x_1, x_2, \dots, x_n , there is defined a probability $F(x_1, x_2, \dots, x_n) = P\{\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n\}$. This function of n arguments is called the *n -dimensional distribution function of the random vector* $(\xi_1, \xi_2, \dots, \xi_n)$.

Later on we will resort to a geometrical illustration and will regard the variables $\xi_1, \xi_2, \dots, \xi_n$ as the coordinates of points in n -dimensional Euclidean space. It is obvious that the position of a point $(\xi_1, \xi_2, \dots, \xi_n)$ depends on chance and that the function $F(x_1, \dots, x_n)$ in such an interpretation yields the probability that the point (ξ_1, \dots, ξ_n) will fall in the n -dimensional parallelepiped $\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n$ with edges parallel to the coordinate axes.

Using the distribution function it is easy to compute the probability that the point $(\xi_1, \xi_2, \dots, \xi_n)$ will be inside the parallelepiped

$$a_i \leq \xi_i < b_i \quad (i=1, 2, \dots, n)$$

where a_i and b_i are arbitrary constants. It is easy to compute that

$$\begin{aligned} P\{a_1 \leq \xi_1 < b_1, a_2 \leq \xi_2 < b_2, \dots, a_n \leq \xi_n < b_n\} = \\ = F(b_1, b_2, \dots, b_n) - \sum_{i=1}^n p_i + \sum_{i < j} p_{ij} \mp \dots \\ \dots + (-1)^n F(a_1, a_2, \dots, a_n) \quad (1) \end{aligned}$$

where $p_{ij\dots k}$ denotes the value of the function $F(c_1, c_2, \dots, c_n)$ for $c_i = a_i, c_j = a_j, \dots, c_k = a_k$ and for the other c_s equal to b_s . We leave the proof of this formula to the reader. We note, for one thing, that $F(x_1, \dots, x_{k-1}, +\infty, x_{k+1}, \dots, x_n)$ gives us the probability that the following system of inequalities will be fulfilled:

$$\xi_1 < x_1, \xi_2 < x_2, \dots, \xi_{k-1} < x_{k-1}, \xi_{k+1} < x_{k+1}, \dots, \xi_n < x_n$$

Since by the extended addition axiom of probabilities

$$\begin{aligned} P\{\xi_1 < x_1, \dots, \xi_{k-1} < x_{k-1}, \xi_{k+1} < x_{k+1}, \dots, \xi_n < x_n\} = \\ = \sum_{s=-\infty}^{\infty} P\{\xi_1 < x_1, \dots, \xi_{k-1} < x_{k-1}, s \leq \xi_k < s+1, \end{aligned}$$

$$\xi_{k+1} < x_{k+1}, \dots, \xi_n < x_n\} = F(x_1, \dots, x_{k-1}, \infty, x_{k+1}, \dots, x_n)$$

it follows that $F(x_1, \dots, x_{k-1}, \infty, x_{k+1}, \dots, x_n)$ is the distribution function of the $(n-1)$ -dimensional random variable $(\xi_1, \dots, \xi_{k-1}, \xi_{k+1}, \dots, \xi_n)$. By further continuing this process we can

determine the k -dimensional distribution functions of any group of k variables $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_k}$ ($i_1 < i_2 < \dots < i_k$) using the formula:

$$F_k(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = \mathbf{P} \{ \xi_{i_1} < x_{i_1}, \dots, \xi_{i_k} < x_{i_k} \} = F(c_1, c_2, \dots, c_n)$$

where $c_s = x_s$ if $s = i_r$ ($1 \leq r \leq k$) and $c_s = +\infty$ in other cases. In particular, the distribution function of the random variable ξ_k is

$$F_k(x) = F(c_1, c_2, \dots, c_n)$$

where all c_i ($i \neq k$) are equal to $+\infty$, and $c_k = x$.

Just as the behaviour of a one-dimensional random variable may be described not only by means of the distribution function but also in other ways, so multidimensional random variables may be defined, say, by means of a nonnegative completely additive set function $\Phi\{E\}$ defined for arbitrary Borel sets of n -dimensional space. We define this function as the probability of the point (ξ_1, \dots, ξ_n) falling in the set E . This method of a probabilistic description of an n -dimensional random variable should be regarded as the most natural one and, theoretically speaking, the most appropriate.

Let us consider some examples.

Example 1. A random vector (ξ_1, \dots, ξ_n) is said to be *uniformly distributed* in the parallelepiped $a_i \leq \xi_i < b_i$ ($1 \leq i \leq n$), if the probability of the point $(\xi_1, \xi_2, \dots, \xi_n)$ falling in any interior domain of this parallelepiped is proportional to its volume and the probability of its falling into the parallelepiped is a sure event.

The distribution function of the sought-for quantity has the form

$$F(x_1, \dots, x_n) = \begin{cases} 0 & \text{if } x_i \leq a_i \text{ for at least one } i \\ \prod_{i=1}^n \frac{c_i - a_i}{b_i - a_i} & \text{where } c_i = x_i \text{ if } a_i \leq x_i \leq b_i \\ & \text{and } c_i = b_i \text{ if } x_i > b_i \end{cases}$$

Example 2. A two-dimensional random variable (ξ_1, ξ_2) is distributed *normally* if its distribution function is

$$F(x, y) = C \int_{-\infty}^x \int_{-\infty}^y e^{-Q(u, v)} du dv$$

Here, $Q(x, y)$ is a positive definite quadratic form of $x-a$ and $y-b$, where a and b are constants.

It is known that a positive definite quadratic form of $x-a$ and $y-b$ may be written in the form

$$Q(x, y) = \frac{(x-a)^2}{2A^2} - r \frac{(x-a)(y-b)}{AB} + \frac{(y-b)^2}{2B^2}$$

where A and B are positive numbers and r , a and b are real numbers, r being subject to the condition $-1 \leq r \leq +1$.

It will readily be seen that for $r^2 \neq 1$ each of the random variables ξ_1 and ξ_2 is subject to a one-dimensional normal law. Indeed,

$$\begin{aligned} F_1(x_1) &= \mathbf{P} \{ \xi_1 < x_1 \} = F(x_1, +\infty) = C \int_{-\infty}^{x_1} \int e^{-Q(x, y)} dx dy = \\ &= C \int_{-\infty}^{x_1} e^{-\frac{(x-a)^2}{2A^2}(1-r^2)} \int e^{-\frac{1}{2} \left[\frac{y-b}{B} - r \frac{x-a}{A} \right]^2} dy dx \end{aligned}$$

Since

$$\int e^{-\frac{1}{2} \left[\frac{y-b}{B} - r \frac{x-a}{A} \right]^2} dy = B \sqrt{2\pi}$$

it follows that

$$F_1(x_1) = BC \sqrt{2\pi} \int_{-\infty}^{x_1} e^{-\frac{(x-a)^2}{2A^2}(1-r^2)} dx \quad (2)$$

The constant C may be expressed in terms of A , B and r . This dependence may be found from the condition $F_1(+\infty) = 1$. We have

$$1 = BC \sqrt{2\pi} \int e^{-\frac{(x-a)^2}{2A^2}(1-r^2)} dx = \frac{ABC \sqrt{2\pi}}{\sqrt{1-r^2}} \int e^{-\frac{z^2}{2}} dz = \frac{2ABC\pi}{\sqrt{1-r^2}}$$

Whence

$$C = \frac{\sqrt{1-r^2}}{2\pi AB}$$

If $r^2 \neq 1$, then we put

$$A = \sigma_1 \sqrt{1-r^2}, \quad B = \sigma_2 \sqrt{1-r^2}$$

In these new notations the two-dimensional normal law takes on the following form:

$$\begin{aligned} F(x_1, x_2) &= \\ &= \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-r^2}} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} e^{-\frac{1}{2(1-r^2)} \left[\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right]} dx dy \end{aligned}$$

The probabilistic meaning of the parameters that enter into this formula will be elucidated in the next chapter.

When $r^2 = 1$ Equation (2) becomes meaningless. In this case, ξ_1 and ξ_2 are connected by a linear relation.

We can establish a number of properties for multidimensional distribution functions just as we did in the one-dimensional case. We will simply state them and leave their proofs to the reader. A distribution function

- (1) is a nondecreasing function of each of its arguments,
- (2) is continuous on the left in each of its arguments,
- (3) satisfies the relations

$$F(+\infty, +\infty, \dots, +\infty) = 1$$

$$\lim_{x_k \rightarrow -\infty} F(x_1, x_2, \dots, x_n) = 0 \quad (1 \leq k \leq n)$$

for arbitrary values of the remaining arguments.

In the one-dimensional case we saw that the enumerated properties are necessary and sufficient for the function $F(x)$ to be a distribution function of some random variable. In the multidimensional case it appears that these properties do not suffice. For the function $F(x_1, \dots, x_n)$ to be a distribution function we have to add the following (in addition to the enumerated three properties):

- (4) for any a_i and b_i ($i=1, 2, \dots, n$) expression (1) is not negative.

That this requirement may not be fulfilled, despite the fact that the function $F(x_1, \dots, x_n)$ has Properties 1 to 3, is seen from the following example. Let

$$F(x, y) = \begin{cases} 0 & \text{if } x \leq 0 \text{ or } x + y \leq 1 \text{ or } y \leq 0 \\ 1 & \text{in the remaining part of the plane} \end{cases}$$

This function satisfies the conditions (1) to (3) but for it

$$F(1, 1) - F\left(1, \frac{1}{2}\right) - F\left(\frac{1}{2}, 1\right) + F\left(\frac{1}{2}, \frac{1}{2}\right) = -1 \quad (3)$$

and, consequently, the fourth condition is not satisfied.

The function $F(x, y)$ cannot be a distribution function because the difference (3) is, according to the relation (1), equal to the probability of point (ξ_1, ξ_2) falling in the rectangle $\frac{1}{2} \leq \xi_1 < 1, \frac{1}{2} \leq \xi_2 < 1$.

If there exists a function $p(x_1, x_2, \dots, x_n)$ such that for any x_1, x_2, \dots, x_n the equation

$$F(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_n} p(z_1, z_2, \dots, z_n) dz_n \dots dz_2 dz_1$$

holds, then this function is called the *probability density function* of the random vector $(\xi_1, \xi_2, \dots, \xi_n)$. It is readily seen that the density function has the following properties:

1. $p(x_1, x_2, \dots, x_n) \geq 0$.
2. The probability of a point $(\xi_1, \xi_2, \dots, \xi_n)$ falling in some domain G is

$$\int_G \dots \int p(x_1, \dots, x_n) dx_n \dots dx_1$$

In particular, if the function $p(x_1, x_2, \dots, x_n)$ is continuous at the point (x_1, \dots, x_n) , then the probability of the point $(\xi_1, \xi_2, \dots, \xi_n)$ falling in the parallelepiped $x_k \leq \xi_k < x_k + dx_k$ ($k=1, 2, \dots, n$) is, to within higher-order infinitesimals,

$$p(x_1, x_2, \dots, x_n) dx_1, dx_2 \dots dx_n$$

Example 3. As an example of an n -dimensional random variable having density, we give a variable uniformly distributed in an n -dimensional domain G . If by V we denote the n -dimensional volume of the domain G , the distribution density will be

$$p(x_1, x_2, \dots, x_n) = \begin{cases} 0 & \text{if } (x_1, x_2, \dots, x_n) \notin G \\ \frac{1}{V} & \text{if } (x_1, x_2, \dots, x_n) \in G \end{cases}$$

Example 4. The density of the two-dimensional normal law is given by the formula

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)} \left[\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right]}$$

We note that the normal density function retains a constant value on the ellipses

$$\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} = \lambda^2 \quad (4)$$

where λ is a constant; for this reason, the ellipses (4) are called *ellipses of equal probabilities*.

Let us find the probability that the point (ξ_1, ξ_2) will fall within the ellipse (4). By the definition of a density function

$$P(\lambda) = \iint_{G(\lambda)} p(x, y) dx dy \quad (5)$$

where $G(\lambda)$ denotes the domain bounded by the ellipse (4). To compute this integral, we introduce the polar coordinates

$$\begin{aligned} x-a &= \rho \cos \theta \\ y-b &= \rho \sin \theta \end{aligned}$$

The integral (5) then takes the form

$$P(\lambda) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_0^{2\pi} \int_0^{\rho^2} e^{-\frac{\rho^2}{2(1-r^2)}} \rho d\rho d\theta$$

where for brevity

$$s^2 = \frac{1}{1-r^2} \left[\frac{\cos^2 \theta}{\sigma_1^2} - 2r \frac{\cos \theta \sin \theta}{\sigma_1 \sigma_2} + \frac{\sin^2 \theta}{\sigma_2^2} \right]$$

Integration with respect to ρ yields

$$P(\lambda) = \frac{1 - e^{-\frac{\lambda^2}{2(1-r^2)}}}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_0^{2\pi} \frac{d\theta}{s^2}$$

Integration with respect to θ may be performed by the rules for integrating trigonometric functions, but this is not necessary since it is automatically carried out by means of probabilistic reasoning. Indeed,

$$P(+\infty) = 1 = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_0^{2\pi} \frac{d\theta}{s^2}$$

Therefore

$$\int_0^{2\pi} \frac{d\theta}{s^2} = 2\pi\sigma_1\sigma_2\sqrt{1-r^2}$$

and, thus,

$$P(\lambda) = 1 - e^{-\frac{\lambda^2}{2(1-r^2)}}$$

The normal distribution plays an exceedingly great role in a variety of applied problems. The distribution of many random variables of practical importance turns out to be subject to the normal distribution law. For instance, the vast experience of artillery firings carried out under a diversity of conditions has shown that shell dispersion on a plane during gunfire from a single gun at a specific target obeys the normal law. In Chapter 8 we will see that the "universality" of the normal law is explained by the fact that any random variable that is the sum of a very large number of independent random variables, each of which exerts only a slight effect on the sum, is distributed almost according to the normal law.

The crucial concept of probability theory—the independence of events—retains its significance for random variables as well. In accord with the definition of the independence of events we can say that the *random variables* $\xi_1, \xi_2, \dots, \xi_n$ are independent if for any group $\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_k}$ ($i_1 < i_2 < \dots < i_k$) of these variables we have the equation

$$\begin{aligned} P\{\xi_{i_1} < x_{i_1}, \xi_{i_2} < x_{i_2}, \dots, \xi_{i_k} < x_{i_k}\} = \\ = P\{\xi_{i_1} < x_{i_1}\} P\{\xi_{i_2} < x_{i_2}\} \dots P\{\xi_{i_k} < x_{i_k}\} \end{aligned}$$

for arbitrary $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ and for any $k (1 \leq k \leq n)$. In particular, the following equation holds for arbitrary x_1, x_2, \dots, x_n :

$$\begin{aligned} \mathbf{P} \{ \xi_1 < x_1, \xi_2 < x_2, \dots, \xi_n < x_n \} &= \\ &= \mathbf{P} \{ \xi_1 < x_1 \} \mathbf{P} \{ \xi_2 < x_2 \} \dots \mathbf{P} \{ \xi_n < x_n \} \end{aligned}$$

or in terms of distribution functions,

$$F(x_1, x_2, \dots, x_n) = F_1(x_1) F_2(x_2) \dots F_n(x_n)$$

where $F_k(x_k)$ denotes the distribution function of the variable ξ_k .

It is readily seen that the converse proposition is also true: if the distribution function $F(x_1, x_2, \dots, x_n)$ of a system of random variables $\xi_1, \xi_2, \dots, \xi_n$ has the form

$$F(x_1, x_2, \dots, x_n) = F_1(x_1) F_2(x_2) \dots F_n(x_n)$$

where the functions $F_k(x_k)$ satisfy the relations

$$F_k(+\infty) = 1 \quad (k = 1, 2, \dots, n)$$

then the variables $\xi_1, \xi_2, \dots, \xi_n$ are independent and the functions $F_1(x_1), F_2(x_2), \dots, F_n(x_n)$ are their distribution functions.

We leave it to the reader to verify this proposition.

If the independent random variables $\xi_1, \xi_2, \dots, \xi_n$ have density functions $p_1(x), p_2(x), \dots, p_n(x)$, then the n -dimensional variable $(\xi_1, \xi_2, \dots, \xi_n)$ has a density function equal to

$$p(x_1, x_2, \dots, x_n) = p_1(x_1) p_2(x_2) \dots p_n(x_n)$$

Example 5. We consider an n -dimensional random variable, the components of which $\xi_1, \xi_2, \dots, \xi_n$ are mutually independent random variables distributed in accordance with the normal laws:

$$F_k(x_k) = \frac{1}{\sigma_k \sqrt{2\pi}} \int_{-\infty}^{x_k} e^{-\frac{(z-a_k)^2}{2\sigma_k^2}} dz$$

In the example at hand, the distribution function is

$$F(x_1, x_2, \dots, x_n) = (2\pi)^{-\frac{n}{2}} \prod_{k=1}^n \sigma_k^{-1} \int_{-\infty}^{x_k} e^{-\frac{(z-a_k)^2}{2\sigma_k^2}} dz$$

The n -dimensional density function of the variable $(\xi_1, \xi_2, \dots, \xi_n)$ is

$$p(x_1, x_2, \dots, x_n) = \frac{(2\pi)^{-\frac{n}{2}}}{\sigma_1 \sigma_2 \dots \sigma_n} e^{-\frac{1}{2} \sum_{k=1}^n \frac{(x_k - a_k)^2}{\sigma_k^2}} \quad (6)$$

For $n=2$ this formula becomes

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{(x_1-a_1)^2}{2\sigma_1^2} - \frac{(x_2-a_2)^2}{2\sigma_2^2}}$$

A comparison of this function with the density of the two-dimensional normal law (Example 4) shows that for the independent random variables ξ_1 and ξ_2 the parameter r is equal to 0.

For $n=3$ formula (6) may be interpreted as the density function of the components ξ_1, ξ_2, ξ_3 of molecular velocity along the coordinate axes (the Maxwell distribution) provided it is assumed that

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \frac{1}{hm}$$

where m is the mass of a molecule and h is a constant.

Example 6. There are two independent random variables ξ and η with distribution functions equal to $F(x)$ and $G(x)$, respectively. Find the probability that η will take on a value less than ξ .

We consider the plane (ξ, η) . For η to be less than ξ , it is necessary that the point (ξ, η) fall in the half-plane $\eta < \xi$. The probability of the simultaneous realization of the inequalities

$$x \leq \xi < x + dx, \quad \eta < x$$

is equal to $G(x) dF(x)$. Since x may have any value from $-\infty$ to $+\infty$, by virtue of the formula of total probability (generalized in obvious fashion) the desired probability is

$$\alpha = \int_{-\infty}^{\infty} G(x) dF(x)$$

In particular, if $G(x) = F(x)$, this probability is

$$\alpha = \int_{-\infty}^{\infty} F(x) dF(x)$$

If the function $F(x)$ is continuous, then

$$\alpha = 0.5$$

There is no such simple result for discrete random variables. This will readily be seen in the case of the random variables ξ and η taking on only two values, 0 and 1, with the probabilities p and $q=1-p$, respectively. In this example it is obvious that

$$\alpha = pq$$

Sec. 24. Functions of Random Variables

The information we have obtained about distribution functions enables us to begin the solution of the following problem: from the distribution function $F(x_1, x_2, \dots, x_n)$ of a collection of random variables $\xi_1, \xi_2, \dots, \xi_n$ determine the distribution function $\Phi(y_1, y_2,$

$\dots, y_k)$ of the variables $\eta_1=f_1(\xi_1, \dots, \xi_n)$, $\eta_2=f_2(\xi_1, \dots, \xi_n)$, \dots , $\eta_k=f_k(\xi_1, \dots, \xi_n)$.

The general solution of this problem is extremely simple but requires an extension of the integral concept. So as not to be drawn aside into purely analytical problems, we confine ourselves to a consideration of the most important special cases: discrete and continuous random variables. In the next section we will give the definition and the principal properties of the Stieltjes integral; there we will give the general form of the most important results of the present section.

Let us first consider the case when the n -dimensional vector (ξ_1, \dots, ξ_n) has a probability density function $p(x_1, x_2, \dots, x_n)$. From the foregoing it is seen that the desired distribution function is defined by the relation

$$\Phi(y_1, y_2, \dots, y_k) = \int_D \dots \int p(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

the region of integration D being determined by the inequalities

$$f_i(x_1, x_2, \dots, x_n) < y_i \quad (i=1, 2, \dots, k)$$

In the case of discrete random variables the solution is obviously given by means of an n -fold sum, which is also extended over the domain D .

We now apply to certain important special cases the general remark that we just made relative to the solution of the general problem that we posed.

The Distribution Function of a Sum. Let it be required to find the distribution function of the sum

$$\eta = \xi_1 + \xi_2 + \dots + \xi_n$$

if $p(x_1, x_2, \dots, x_n)$ is the probability density function of the vector $(\xi_1, \xi_2, \dots, \xi_n)$. The desired function is equal to the probability of the point $(\xi_1, \xi_2, \dots, \xi_n)$ falling in the half-space $\xi_1 + \xi_2 + \dots + \xi_n < x$ and consequently

$$\Phi(x) = \int \dots \int_{\sum x_k < x} p(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n$$

We consider in more detail the case of $n=2$. The preceding formula now takes the form

$$\Phi(x) = \int \int_{x_1 + x_2 < x} p(x_1, x_2) dx_1 dx_2 = \int_{-\infty}^{x-x_2} \int p(x_1, x_2) dx_1 dx_2 \quad (1)$$

If the variables ξ_1 and ξ_2 are independent, then $p(x_1, x_2) = p_1(x_1)p_2(x_2)$ and Equation (1) may be written in the following

form:

$$\begin{aligned}\Phi(x) &= \int dx_1 \int_{-\infty}^{x-x_1} p_1(x_1) p_2(x_2) dx_2 = \int dx_1 \int_{-\infty}^x p_1(x_1) p_2(z-x_1) dz = \\ &= \int_{-\infty}^x dz \left\{ \int p_1(x_1) p_2(z-x_1) dx_1 \right\} \quad (2)\end{aligned}$$

In the general case, formula (1) yields

$$\Phi(x) = \int_{-\infty}^x dx_1 \int p(z, x_1 - z) dz \quad (3)$$

The last equations prove that if a multidimensional distribution of summands has a probability density function, then their sum also has a density function. This density, in the case of independent summands, may be written as

$$p(x) = \int p_1(x-z) p_2(z) dz \quad (4)$$

Let us consider some examples.

Example 1. Let ξ_1 and ξ_2 be independent and uniformly distributed in the interval (a, b) . Find the density function of the sum $\eta = \xi_1 + \xi_2$.

The probability density functions of ξ_1 and ξ_2 are equal to

$$p_1(x) = p_2(x) = \begin{cases} 0 & \text{if } x \leq a \text{ or } x > b \\ \frac{1}{b-a} & \text{if } a < x \leq b \end{cases}$$

Using formula (4) we find that

$$p_\eta(x) = \int_a^b p_1(z) p_2(x-z) dz = \frac{1}{b-a} \int_a^b p_2(x-z) dz$$

From the fact that for $x < 2a$

$$x-z < 2a-z < a$$

and for $x > 2b$

$$x-z > 2b-z > b$$

we conclude that for $x < 2a$ and $x > 2b$

$$p_\eta(x) = 0$$

Now let $2a < x < 2b$. The integrand is different from zero for only those values of z which satisfy the inequalities

$$a < x - z < b$$

or, which is the same thing, the inequalities

$$x - b < z < x - a$$

Since $x > 2a$, it follows that $x - a > a$. Obviously, $x - a \leq b$ for $x \leq a + b$. Hence, if $2a < x \leq a + b$, then it follows that

$$p_{\eta}(x) = \int_a^{x-a} \frac{dz}{(b-a)^2} = \frac{x-2a}{(b-a)^2}$$

In exactly the same way, when $a + b < x \leq 2b$

$$p_{\eta}(x) = \int_{x-b}^b \frac{dz}{(b-a)^2} = \frac{2b-x}{(b-a)^2}$$

Collecting together the results obtained, we find

$$p_{\eta}(x) = \begin{cases} 0 & \text{for } x \leq 2a \text{ and } x > 2b \\ \frac{x-2a}{(b-a)^2} & \text{for } 2a < x \leq a+b \\ \frac{2b-x}{(b-a)^2} & \text{for } a+b < x \leq 2b \end{cases} \quad (5)$$

The function $p_{\eta}(x)$ is called the *Simpson distribution law*.

In the example considered, the computations are greatly simplified if we take advantage of geometrical reasoning. As usual, depict ξ_1 and ξ_2 as rectangular coordinates in the plane. Then the probability of the inequality $\xi_1 + \xi_2 < x$ for $2a < x \leq a + b$ is equal to the probability of falling in the doubly cross-hatched right triangle (Fig. 15). This probability is readily found to be

$$F_{\eta}(x) = \frac{(x-2a)^2}{2(a-b)^2}$$

For $a + b < x \leq 2b$, the probability of the inequality $\xi_1 + \xi_2 < x$ is equal to the probability of falling in the entire shaded figure. This probability is

$$F_{\eta}(x) = 1 - \frac{(2b-x)^2}{2(b-a)^2}$$

Differentiation with respect to x leads us to the formula (5).

In connection with the example we have considered it is interesting to note the following.

General problems of geometry led N. I. Lobachevsky to the necessity of solving the following problem: given a group of n mutually independent random variables $\xi_1, \xi_2, \dots, \xi_n$ (errors of observation), find the probability distribution of their arithmetic mean.

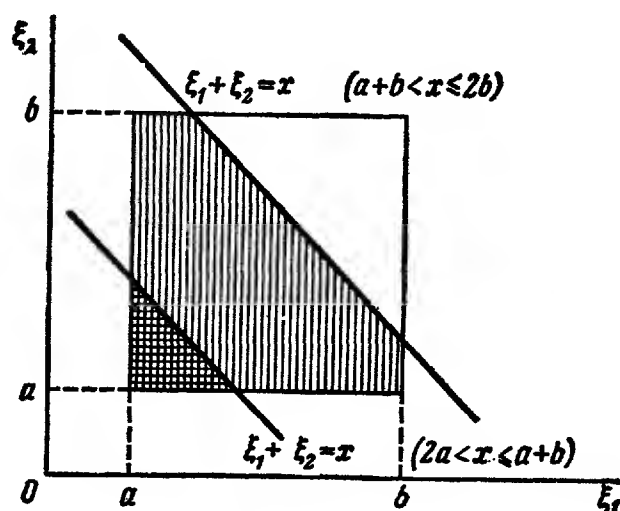


Fig. 15

He solved this problem only for the case when all errors are *uniformly* distributed in the interval $(-1, 1)$, and it was found that the probability of an arithmetic-mean error lying within the limits from $-x$ to x is

$$P_n(x) = 1 - \frac{1}{2^{n-1}} \sum (-1)^r \frac{[n-nx-2r]^r}{r! (n-r)!}$$

where the summation extends over all integral r from $r=0$ to $r = \left\lfloor \frac{n-nx}{2} \right\rfloor$.

Example 2. A two-dimensional random variable (ξ_1, ξ_2) is distributed in accordance with the normal law:

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-r^2}} \times \\ \times \exp \left\{ -\frac{1}{2(1-r^2)} \left(\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right) \right\}$$

Find the distribution function of the sum $\eta = \xi_1 + \xi_2$.

According to the formula (3)

$$p_\eta(x) = \frac{1}{2\pi\sigma_1\sigma_2 \sqrt{1-r^2}} \times \\ \times \int \exp \left\{ -\frac{1}{2(1-r^2)} \left(\frac{(z-a)^2}{\sigma_1^2} - 2r \frac{(z-a)(x-z-b)}{\sigma_1\sigma_2} + \frac{(x-z-b)^2}{\sigma_2^2} \right) \right\} dz$$

For the sake of brevity, denote $x-a-b$ by v and $z-a$ by u , then

$$p_{\eta}(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \int \exp \left\{ -\frac{1}{2(1-r^2)} \left(\frac{u^2}{\sigma_1^2} - 2r \frac{u(v-u)}{\sigma_1\sigma_2} + \frac{(v-u)^2}{\sigma_2^2} \right) \right\} du$$

Since

$$\begin{aligned} \frac{u^2}{\sigma_1^2} - 2r \frac{u(v-u)}{\sigma_1\sigma_2} + \frac{(v-u)^2}{\sigma_2^2} &= u^2 \frac{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}{\sigma_1^2\sigma_2^2} - 2uv \frac{\sigma_1 + r\sigma_2}{\sigma_1\sigma_2^2} + \frac{v^2}{\sigma_2^2} = \\ &= \left[u \frac{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}}{\sigma_1\sigma_2} - \frac{v}{\sigma_2} \frac{\sigma_1 + r\sigma_2}{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}} \right]^2 + \\ &\quad + \frac{v^2}{\sigma_2^2} \left(1 - \frac{(\sigma_1 + r\sigma_2)^2}{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2} \right) = \\ &= \left[u \frac{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}}{\sigma_1\sigma_2} - \frac{v}{\sigma_2} \frac{\sigma_1 + r\sigma_2}{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}} \right]^2 + \frac{v^2(1-r^2)}{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2} \end{aligned}$$

it follows that by introducing the notation

$$t = \frac{1}{\sqrt{1-r^2}} \left[u \frac{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}}{\sigma_1\sigma_2} - \frac{v}{\sigma_2} \frac{\sigma_1 + r\sigma_2}{\sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}} \right]$$

we will reduce the expression for $p_{\eta}(x)$ to the form

$$p_{\eta}(x) = \frac{\exp \left\{ -\frac{v^2}{2(\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2)} \right\}}{2\pi \sqrt{\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2}} \int e^{-\frac{t^2}{2}} dt$$

Since

$$v = x - a - b \quad \text{and} \quad \int e^{-\frac{t^2}{2}} dt = \sqrt{2\pi}$$

it follows that

$$p_{\eta}(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2)}} e^{-\frac{(x-a-b)^2}{2(\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2)}}$$

In particular, if the random variables ξ_1 and ξ_2 are independent, then $r=0$, and the formula for $p_{\eta}(x)$ takes on the form

$$p_{\eta}(x) = \frac{1}{\sqrt{2\pi(\sigma_1^2 + \sigma_2^2)}} e^{-\frac{(x-a-b)^2}{2(\sigma_1^2 + \sigma_2^2)}}$$

As a result of this substitution,

$$\Phi(y) = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \dots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{y\sqrt{n}} \frac{1}{(\sqrt{2\pi})^n} e^{-\frac{\rho^2}{2}} \rho^{n-1} D(\theta_1 \dots \theta_{n-1}) d\rho d\theta_{n-1} \dots$$

$$\dots d\theta_1 = C_n \int_0^{y\sqrt{n}} e^{-\frac{\rho^2}{2}} \rho^{n-1} d\rho$$

where the constant

$$C_n = \frac{1}{(\sqrt{2\pi})^n} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \dots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} D(\theta_1 \dots \theta_{n-1}) d\theta_{n-1} \dots d\theta_1$$

is dependent solely on n .

This constant is readily calculable by using the equation

$$\Phi(+\infty) = 1 = C_n \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho^{n-1} d\rho = C_n \Gamma\left(\frac{n}{2}\right) \cdot 2^{\frac{n}{2}-1}$$

From this we find

$$\Phi(y) = \frac{1}{2^{\frac{n}{2}-1} \Gamma\left(\frac{n}{2}\right)} \int_0^{y\sqrt{n}} \rho^{n-1} e^{-\frac{\rho^2}{2}} d\rho$$

The density function of the random variable ξ for $y \geq 0$ is

$$\varphi(y) = \frac{\sqrt{2n}}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{y\sqrt{n}}{\sqrt{2}}\right)^{n-1} e^{-\frac{ny^2}{2}} \quad (6)$$

Whence, in particular, for $n=1$, we naturally get a density function equal to twice the density of the initial normal law:

$$\varphi(y) = \sqrt{\frac{2}{\pi}} e^{-\frac{y^2}{2}} \quad (y \geq 0)$$

For $n=3$ we get the familiar Maxwell law

$$\varphi(y) = \frac{3\sqrt{6}}{\sqrt{\pi}} y^2 e^{-\frac{3y^2}{2}}$$

It is easy to derive the density function of the variable χ^2 either by computations similar to those carried out or directly from formula (6). When $x \geq 0$, this density function is

$$\rho_n(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} \quad (6')$$

The following is a table of variables associated with χ^2 and frequently used in practical problems:

Variable	Density function for $x \geq 0$
$\chi^2 = \frac{1}{\sigma^2} \sum_{k=1}^n (\xi_k - a)^2$	$\frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$
$\frac{1}{n} \chi^2 = \frac{1}{n\sigma^2} \sum_{k=1}^n (\xi_k - a)^2$	$\frac{\left(\frac{n}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{nx}{2}}$
$\chi = \sqrt{\frac{1}{\sigma^2} \sum_{k=1}^n (\xi_k - a)^2}$	$\frac{2}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} x^{n-1} e^{-\frac{x^2}{2}}$
$\zeta = \frac{\chi}{\sqrt{n}} = \sqrt{\frac{1}{n\sigma^2} \sum_{k=1}^n (\xi_k - a)^2}$	$\frac{\sqrt{2n}}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{x \sqrt{n}}{\sqrt{2}}\right)^{n-1} e^{-\frac{nx^2}{2}}$

Example 4. The Distribution Function of a Quotient. Let the probability density function of a variable (ξ, η) be $p(x, y)$. It is required to find the distribution function of the quotient $\zeta = \frac{\xi}{\eta}$.

By definition,

$$F_{\zeta}(x) = \mathbf{P}\left\{\frac{\xi}{\eta} < x\right\}$$

If ξ and η are regarded as rectangular Cartesian coordinates of a point in a plane, then $F_{\zeta}(x)$ is equal to the probability that the point (ξ, η) will fall in a region whose point coordinates satisfy the inequality $\frac{\xi}{\eta} < x$. This region is shaded in Fig. 16.

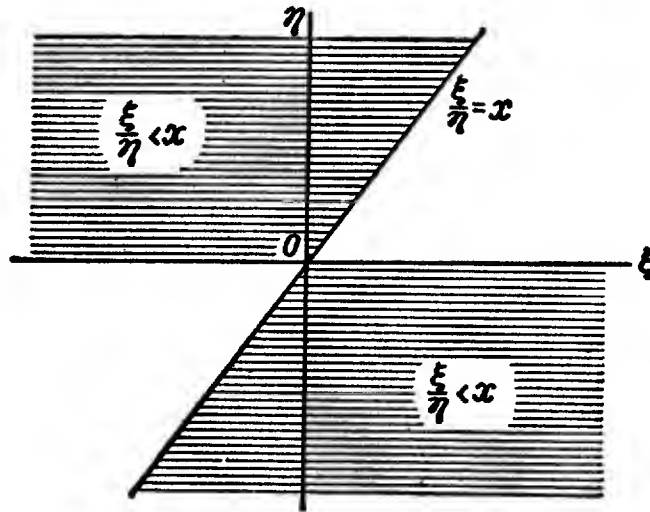


Fig. 16

According to the general formula, the sought-for probability is

$$F_{\zeta}(x) = \int_0^{\infty} \int_{-\infty}^{zx} p(y, z) dy dz + \int_{-\infty}^0 \int_{zx}^{\infty} p(y, z) dy dz \quad (7)$$

Whence it follows that if ξ and η are independent, and $p_1(x)$ and $p_2(x)$ are their density functions, then

$$F_{\zeta}(x) = \int_0^{\infty} F_1(xz) p_2(z) dz + \int_{-\infty}^0 [1 - F_1(xz)] p_2(z) dz \quad (7')$$

Differentiating (7) we find

$$p_{\zeta}(x) = \int_0^{\infty} z p(zx, z) dz - \int_{-\infty}^0 z p(zx, z) dz = \int_{-\infty}^{\infty} |z| p(zx, z) dz \quad (8)$$

In particular, if ξ and η are independent, then

$$p_{\zeta}(x) = \int_0^{\infty} z p_1(zx) p_2(z) dz - \int_{-\infty}^0 z p_1(zx) p_2(z) dz \quad (8')$$

Example 5. The random variable (ξ, η) is distributed according to the normal law:

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{x^2}{\sigma_1^2} - 2r \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right] \right\}$$

Find the distribution function of the quotient $\zeta = \frac{\xi}{\eta}$.

From formula (8)

$$\begin{aligned} p_\zeta(x) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \\ &\times \left[\int_0^\infty - \int_{-\infty}^0 \right] z \exp \left\{ -\frac{z^2}{2(1-r^2)} \left[\frac{\sigma_2^2 x^2 - 2r\sigma_1\sigma_2 x + \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right] \right\} dz = \\ &= \frac{1}{\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int_0^\infty z \exp \left\{ -\frac{z^2}{2(1-r^2)} \cdot \frac{\sigma_2^2 x^2 - 2r\sigma_1\sigma_2 x + \sigma_1^2}{\sigma_1^2 \sigma_2^2} \right\} dz \end{aligned}$$

Perform a substitution in the integral by putting

$$u = \frac{z^2}{2(1-r^2)} \frac{\sigma_2^2 x^2 - 2r\sigma_1\sigma_2 x + \sigma_1^2}{\sigma_1^2 \sigma_2^2}$$

The expression for $p_\zeta(x)$ will then become

$$p_\zeta(x) = \frac{\sigma_1\sigma_2\sqrt{1-r^2}}{\pi(\sigma_2^2 x^2 - 2r\sigma_1\sigma_2 x + \sigma_1^2)} \int_0^\infty e^{-u} du = \frac{\sigma_1\sigma_2\sqrt{1-r^2}}{\pi(\sigma_2^2 x^2 - 2r\sigma_1\sigma_2 x + \sigma_1^2)}$$

In particular, if the variables ξ and η are independent, then

$$p_\zeta(x) = \frac{\sigma_1\sigma_2}{\pi(\sigma_1^2 + \sigma_2^2 x^2)}$$

The density function of the variable ζ is called *Cauchy's law*.

Example 6. Student's Distribution. Find the distribution function of the quotient $\zeta = \xi/\eta$, where ξ and η are independent variables, ξ being distributed according to the normal law:

$$p_\xi(x) = \sqrt{\frac{n}{2\pi}} e^{-\frac{nx^2}{2}}$$

and $\eta = \frac{\chi}{\sqrt{n}}$ (see Example 3), so that

$$p_\eta(x) = \frac{\sqrt{2n}}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{y\sqrt{n}}{\sqrt{2}} \right)^{n-1} e^{-\frac{ny^2}{2}}$$

According to the formula (8')

$$\begin{aligned} p_{\zeta}(x) &= \int_0^{\infty} z \sqrt{\frac{n}{2\pi}} e^{-\frac{nz^2x^2}{2}} \frac{\sqrt{2n}}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{z\sqrt{n}}{\sqrt{2}}\right)^{n-1} e^{-\frac{nz^2}{2}} dz = \\ &= \frac{1}{\sqrt{\pi} \Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} \left(\frac{z\sqrt{n}}{\sqrt{2}}\right)^{n-1} e^{-\frac{nz^2}{2}(x^2+1)} nz dz \end{aligned}$$

Substituting

$$u = \frac{nz^2}{2} (x^2 + 1)$$

we find that

$$p_{\zeta}(x) = \frac{(x^2 + 1)^{-\frac{n+1}{2}}}{\sqrt{\pi} \Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} u^{\frac{n-1}{2}} e^{-u} du = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n}{2}\right)} (x^2 + 1)^{-\frac{n+1}{2}}$$

The probability density function

$$p_{\zeta}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{n}{2}\right)} (1 + x^2)^{-\frac{n+1}{2}}$$

is called *Student's law**.

For $n=1$, Student's law becomes Cauchy's law.

Example 7. Rotation of the Coordinate Axes. Given the distribution function of a two-dimensional random variable (ξ, η) , find the distribution function of the variables

$$\begin{aligned} \xi' &= \xi \cos \alpha + \eta \sin \alpha, \quad \eta' = \\ &= -\xi \sin \alpha + \eta \cos \alpha \end{aligned} \quad (9)$$

Denote by $F(x, y)$ and $\Phi(x, y)$ the distribution functions of the variables (ξ, η) and (ξ', η') . If we regard (ξ, η) and (ξ', η') as the rectangular Cartesian coordinates of a point in the plane, then it will be easy to see that the coordinate system $O\xi'\eta'$ is obtained from the system $O\xi\eta$ by rotation of the axes through the angle α . We confine ourselves to the case $0 < \alpha < \frac{\pi}{2}$, leaving it to the reader to derive analogous formulas for the remaining values of α .

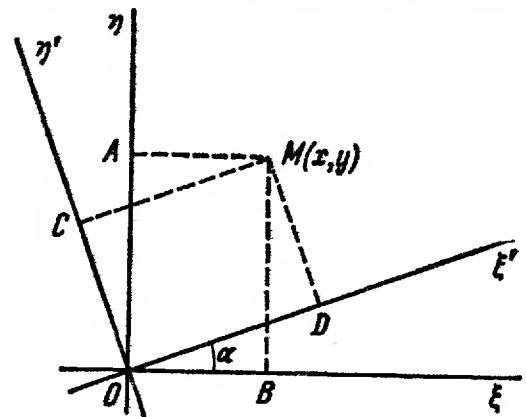


Fig. 17

* Student was the pseudonym of the English statistician W. L. Gosset, who first discovered this law in empirical fashion.

Let the coordinates of the point M in the $O\xi\eta$ system be x and y , and in the $O\xi'\eta'$ system, x' and y' . Then the function $F(x, y)$ is equal to the probability of the point (ξ, η) falling in the right angle bounded by the half-lines AB and BM (Fig. 17), and the function $\Phi(x', y')$ is equal to the probability of the point (ξ, η) falling in the angle bounded by the half-lines CM and DM . The equations of the straight lines CM and DM in the coordinate system $O\xi\eta$ is of the form: for CM ,

$$\eta = (\xi - x) \tan \alpha + y$$

and for DM

$$\eta = -(\xi - x) \cot \alpha + y$$

Since (x, y) and (x', y') are connected by the equations

$$x' = x \cos \alpha + y \sin \alpha, \quad y' = -x \sin \alpha + y \cos \alpha$$

these equations may be written in a different form:

$$\eta = \xi \tan \alpha + \frac{y'}{\cos \alpha}$$

$$\eta = -\xi \cot \alpha + \frac{x'}{\sin \alpha}$$

By virtue of what has already been said,

$$\Phi(x', y') = \iint p(\xi, \eta) d\eta d\xi$$

The integral is extended over the interior part of the angle CMD . It is easy to see that

$$\begin{aligned} \Phi(x', y') = \int_{-\infty}^x \int_{-\infty}^{\xi \tan \alpha + \frac{y'}{\cos \alpha}} p(\xi, \eta) d\eta d\xi + \\ + \int_x^{\infty} \int_{-\infty}^{-\xi \cot \alpha + \frac{x'}{\sin \alpha}} p(\xi, \eta) d\eta d\xi \end{aligned}$$

Differentiating this equation with respect to x' and y' , we find

$$\begin{aligned} \pi(x', y') = \frac{\partial \Phi(x', y')}{\partial x' \partial y'} = p(x, y) = \\ = p(x' \cos \alpha - y' \sin \alpha, x' \sin \alpha + y' \cos \alpha) \quad (10) \end{aligned}$$

Example 8. The two-dimensional random variable (ξ, η) is distributed according to the normal law:

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{x^2}{\sigma_1^2} - 2r \frac{xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2} \right] \right\}$$

Find the density function of the random variables

$$\xi' = \xi \cos \alpha + \eta \sin \alpha, \quad \eta' = -\xi \sin \alpha + \eta \cos \alpha$$

By (10) we have

$$\begin{aligned} \pi(x', y') &= p(x' \cos \alpha - y' \sin \alpha, x' \sin \alpha + y' \cos \alpha) = \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} [Ax'^2 - 2Bx'y' + Cy'^2] \right\} \end{aligned}$$

where

$$\begin{aligned} A &= \frac{\cos^2 \alpha}{\sigma_1^2} - 2r \frac{\cos \alpha \sin \alpha}{\sigma_1 \sigma_2} + \frac{\sin^2 \alpha}{\sigma_2^2} \\ B &= \frac{\cos \alpha \sin \alpha}{\sigma_1^2} - r \frac{\sin^2 \alpha - \cos^2 \alpha}{\sigma_1 \sigma_2} - \frac{\cos \alpha \sin \alpha}{\sigma_2^2} \\ C &= \frac{\sin^2 \alpha}{\sigma_1^2} + 2r \frac{\cos \alpha \sin \alpha}{\sigma_1 \sigma_2} + \frac{\cos^2 \alpha}{\sigma_2^2} \end{aligned}$$

From the formula obtained we conclude that rotation of the coordinate axes transforms a normal distribution into a normal distribution.

It will be noted that if the angle α is chosen so that

$$\tan 2\alpha = \frac{2r\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

then $B=0$ and

$$\pi(x', y') = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{Ax'^2}{2(1-r^2)} - \frac{Cy'^2}{2(1-r^2)}}$$

This equation implies that any normally distributed two-dimensional random variable may, by rotation of the coordinate axes, be reduced to a system of two normally distributed *independent* random variables. This result can be extended to n -dimensional random variables.

It is possible to prove a stronger proposition that exhaustively describes a normal probability distribution. Let there be a *nondegenerate* (i.e., not concentrated on a single straight line) probability distribution in the plane. For this distribution to be normal, it is necessary and sufficient that it be possible, in two different ways, to choose in the plane the coordinate axes $O\xi_1\xi_2$ and $O\eta_1\eta_2$ such that the coordinates ξ_1 and ξ_2 (just as η_1 and η_2), regarded as random variables with a given probability distribution, should be independent.

Sec. 25. The Stieltjes Integral

The Stieltjes integral will be made substantial use of in what follows, and so to facilitate the study of subsequent sections we give here a definition and the basic properties of the Stieltjes integral without dwelling on proofs.

Suppose in an interval (a, b) we have defined the function $f(x)$ and a nondecreasing function $F(x)$ of bounded variation. For the sake of definiteness we will assume here that the function $F(x)$ is continuous on the left. If a and b are finite, we partition the interval (a, b) into a finite number of subintervals (x_i, x_{i+1}) by means of the points $a = x_0 < x_1 < x_2 \dots < x_n = b$ and form the sum

$$\sum_{i=1}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})]$$

where \tilde{x}_i is an arbitrary number chosen in the interval (x_{i-1}, x_i) . We now increase the number of partition points and at the same time let the length of the largest subinterval approach zero. If in this process the foregoing sum tends to a definite limit

$$J = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] \quad (1)$$

then this limit is called the *Stieltjes integral* of the function $f(x)$ with respect to the integrating function $F(x)$ and is denoted by the symbol

$$J = \int_a^b f(x) dF(x) \quad (2)$$

When the interval of integration is infinite, the improper Stieltjes integral is defined in the usual way: the integral is considered over an arbitrary finite interval (a, b) ; the quantities a and b are made to approach $-\infty$ and $+\infty$ in arbitrary fashion; if there exists a limit

$$\lim_{\substack{a \rightarrow -\infty \\ b \rightarrow \infty}} \int_a^b f(x) dF(x)$$

then this limit is called the Stieltjes integral of the function $f(x)$ with respect to the function $F(x)$ in the interval $(-\infty, \infty)$ and is denoted by

$$\int f(x) dF(x)$$

It may be proved that if the function $f(x)$ is continuous and bounded, then the limit of the sum (1) exists in the case of both finite and infinite limits of integration.

In certain cases the Stieltjes integral exists for unbounded functions $f(x)$ as well. Such integrals are of considerable interest to probability theory (expectation, variance, moments, and others).

Everywhere henceforth we will consider that the integral of a function $f(x)$ exists when and only when there exists an integral of $|f(x)|$ with respect to the same integrating function $F(x)$.

For the purposes of probability theory it is important to extend the definition of the Stieltjes integral to the case when the function $f(x)$ may have a finite or countable set of discontinuity points. It may be proved* that any bounded function having a finite or countable set of points of discontinuity, in particular, any function of bounded variation, is integrable with respect to any integrating function of bounded variation. It is then necessary to modify somewhat the definition of the Stieltjes integral; namely, when forming the limit (1) it is necessary to consider only those sequences of subdivisions of the interval of integration such that each point of discontinuity of $f(x)$ is one of the partition points of all subdivisions with the exception, perhaps, of a finite number of them.

It should be noted that when establishing the limits of integration it is important to indicate whether one or another end point of the interval of integration is included or not. Indeed, from the definition of the Stieltjes integral we obtain the following equation (the symbol $a-0$ implies that a is included in the interval of integration and the symbol $a+0$ that a is excluded):

$$\begin{aligned} \int_{a-0}^b f(x) dF(x) &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] = \\ &= \lim_{n \rightarrow \infty} \sum_{i=2}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] + \lim_{x_1 \rightarrow x_0=a} f(\tilde{x}) [F(x_1) - F(x_0)] = \\ &= \int_{a+0}^b f(x) dF(x) + f(a) [F(a+0) - F(a)] \end{aligned}$$

Thus, if $f(a) \neq 0$ and the function $F(x)$ has a jump at $x=a$, then

$$\int_{a-0}^b f(x) dF(x) - \int_{a+0}^b f(x) dF(x) = f(a) [F(a+0) - F(a-0)]$$

This means that the Stieltjes integral, extended over an interval that reduces to a single point, can yield a result different from zero. From now on we shall agree that unless otherwise stated the right end point of the interval will be excluded and the left end point will be included in the interval of integration. This condition permits us to write the following equation:

$$\int_a^b dF(x) = F(b) - F(a)$$

* See V. I. Glivenko, *The Stieltjes Integral*, p. 116 (in Russian).

Indeed, by definition,

$$\int_a^b dF(x) = \lim_{n \rightarrow \infty} \sum_{i=1}^n [F(x_i) - F(x_{i-1})] = \lim_{n \rightarrow \infty} [F(x_n) - F(x_0)] = F(b) - F(a)$$

(recall that by definition $F(x)$ is continuous on the left and, hence, for it $F(b) = \lim_{\varepsilon \rightarrow 0} F(b - \varepsilon)$).

In particular, if $F(x)$ is the distribution function of a random variable ξ , then

$$\begin{aligned} \int_a^b dF(x) &= F(b) - F(a) = \mathbf{P}\{a \leq \xi < b\} \\ \int_{-\infty}^b dF(x) &= F(b) = \mathbf{P}\{\xi < b\} \end{aligned}$$

If $F(x)$ has a derivative of which it is an integral, then from the fact that by the formula of finite increments

$$F(x_i) - F(x_{i-1}) = p(\tilde{x}_i)(x_i - x_{i-1})$$

where $x_{i-1} < \tilde{x}_i < x_i$ there follows the equation

$$\begin{aligned} \int_a^b f(x) dF(x) &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] = \\ &= \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\tilde{x}_i) p(\tilde{x}_i)(x_i - x_{i-1}) = \int_a^b f(x) p(x) dx \end{aligned}$$

We see that in this case the Stieltjes integral reduces to an ordinary integral.

If $F(x)$ has a jump at the point $x=c$, then by selecting the subdivisions so that for certain values of the subscript $x_k < c < x_{k+1}$ we have

$$\begin{aligned} \int_a^b f(x) dF(x) &= \lim_{n \rightarrow \infty} \sum_{i=1}^k f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] + \\ &+ f(c) [F(x_{k+1}) - F(x_k)] + \lim_{n \rightarrow \infty} \sum_{i=k+2}^n f(\tilde{x}_i) [F(x_i) - F(x_{i-1})] = \\ &= \int_a^c f(x) dF(x) + \int_{c+0}^b f(x) dF(x) + f(c) [F(c+0) - F(c-0)] \end{aligned}$$

In particular, if the value of the function $F(x)$ changes only at the points $c_1, c_2, \dots, c_n, \dots$, then

$$\int_a^b f(x) dF(x) = \sum_{n=1}^{\infty} f(c_n) [F(c_n + 0) - F(c_n - 0)]$$

and the Stieltjes integral reduces to a series.

We will now enumerate the principal properties of the Stieltjes integral that we will need in what follows. The reader will find no difficulty in providing the proofs of these properties by proceeding from the definition of the Stieltjes integral and taking advantage of the reasoning used in the theory of the ordinary integral.

1. For $a < c_1 < c_2 < \dots < c_n < b$

$$\int_a^b f(x) dF(x) = \sum_{i=0}^n \int_{c_i}^{c_{i+1}} f(x) dF(x) \quad [a = c_0, \quad b = c_{n+1}]$$

2. A constant factor may be removed from under the integral sign:

$$\int_a^b cf(x) dF(x) = c \int_a^b f(x) dF(x)$$

3. The integral of a sum of functions is equal to the sum of their integrals:

$$\int_a^b \sum_{i=1}^n f_i(x) dF(x) = \sum_{i=1}^n \int_a^b f_i(x) dF(x)$$

4. If $f(x) \geq 0$ and $b > a$, then

$$\int_a^b f(x) dF(x) \geq 0$$

5. If $F_1(x)$ and $F_2(x)$ are monotone functions of bounded variation and c_1 and c_2 are arbitrary constants, then

$$\int_a^b f(x) d[c_1 F_1(x) + c_2 F_2(x)] = c_1 \int_a^b f(x) dF_1(x) + c_2 \int_a^b f(x) dF_2(x)$$

6. If $F(x) = \int_c^x g(u) dG(u)$, where c is a constant, $g(u)$ is a continuous function and $G(u)$ is a nondecreasing function of bounded variation, then

$$\int_a^b f(x) dF(x) = \int_a^b f(x) g(x) dG(x)$$

Employing the concept of the Stieltjes integral, we can write general formulas for the distribution function of the sum of two independent random variables, ξ_1 and ξ_2 :

$$F(x) = \int F_1(x-z) dF_2(z) = \int F_2(x-z) dF_1(z)$$

and also for their quotient $\frac{\xi_1}{\xi_2}$:

$$F(x) = \int_0^{\infty} F_1(xz) dF_2(z) + \int_{-\infty}^0 [1 - F_1(xz)] dF_2(z)$$

on the assumption that $P\{\xi_2 = 0\} = 0$.

EXERCISES

1. Prove that if $F(x)$ is a distribution function, then for any $h \neq 0$ the functions

$$\Phi(x) = \frac{1}{h} \int_x^{x+h} F(x) dx, \quad \Psi(x) = \frac{1}{2h} \int_{x-h}^{x+h} F(x) dx$$

are also distribution functions.

2. A random variable ξ has $F(x)$ as its distribution function ($p(x)$ is the density function). Find the distribution function (density function) of the random variable:

- (a) $\eta = a\xi + b$, a and b are real numbers;
- (b) $\eta = \xi^{-1}$ ($P\{\xi = 0\} = 0$);
- (c) $\eta = \tan \xi$;
- (d) $\eta = \cos \xi$;
- (e) $\eta = f(\xi)$, where $f(x)$ is a continuous monotone function without intervals of constancy.

3. From the point $(0, a)$ draw a straight line at an angle φ to the y -axis. Find the distribution function for the abscissa of the point of intersection of this line with the x -axis if

(a) the angle φ is uniformly distributed in the interval $\left(0, \frac{\pi}{2}\right)$;

(b) the angle φ is uniformly distributed in the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

4. A point is thrown at random on the circumference of a circle of radius R with centre at the coordinate origin [in other words, the polar angle of the point of impact is uniformly distributed in the interval $(-\pi, \pi)$]. Find the density function for:

- (a) the abscissa of the point of impact;
- (b) the length of the chord connecting the point of impact with the point $(-R, 0)$.

5. A point is thrown at random on the segment of the ordinate axis between points $(0, 0)$ and $(0, R)$ [that is, the ordinate of the point is uniformly distributed in the interval $(0, R)$]. Through the point of impact draw a chord of the circle $x^2 + y^2 = R^2$ perpendicular to the y -axis. Find the distribution of length of the chord.

6. The diameter of a circle is measured approximately. Considering that it is uniformly distributed in the interval (a, b) , find the distribution of the area of the circle.

7. The density function of a random variable ξ is given by the equation

$$p(x) = \frac{a}{e^{-x} + e^x}$$

Find:

- (a) the constant a ;
- (b) the probability that in two independent observations ξ will take on values less than 1.

8. The distribution function of a random vector (ξ, η) is of the form:

- (a) $F(x, y) = F_1(x)F_2(y) + F_3(x)$;
- (b) $F(x, y) = F_1(x)F_2(y) + F_3(x) + F_4(y)$.

Can the functions $F_3(x)$ and $F_4(x)$ be arbitrary? Are the components of the vector (ξ, η) dependent or independent?

9. Two points are dropped at random on the interval $(0, a)$ [that is, their abscissas are uniformly distributed on the interval $(0, a)$]. Find the distribution function of the distance between them.

10. A total of n points are dropped on the interval $(0, a)$. Assuming that the points have been dispersed at random [that is, each of them is situated irrespective of the others and is distributed uniformly on $(0, a)$], find:

- (a) the density function of the abscissa of the k th point on the left;
- (b) the joint density function of the abscissas of the k th and m th points on the left ($k < m$).

11. A total of n independent trials are performed on a random variable ξ having a continuous distribution function, as a result of which the following values of the variable ξ were observed: x_1, x_2, \dots, x_n . Find the distribution functions of the random variables:

- (a) $\eta_n = \max(x_1, x_2, \dots, x_n)$;
- (b) $\zeta_n = \min(x_1, x_2, \dots, x_n)$;
- (c) the k th largest observation result;
- (d) the joint distribution of the k th and m th largest observed values.

12. The distribution function of the random vector $(\xi_1, \xi_2, \dots, \xi_n)$ is $F(x_1, x_2, \dots, x_n)$. As the result of a trial the components of the vector take on the values (z_1, z_2, \dots, z_n) . Find the distribution function of the random variable:

- (a) $\eta_n = \max(z_1, z_2, \dots, z_n)$;
- (b) $\zeta_n = \min(z_1, z_2, \dots, z_n)$.

13. The random variable ξ has a continuous distribution function $F(x)$. How is the random variable $\eta = F(\xi)$ distributed?

14. The random variables ξ and η are independent; their density functions are defined by the equations

$$\begin{aligned} p_\xi(x) &= p_\eta(x) = 0 & \text{for } x \leq 0 \\ p_\xi(x) &= c_1 x^\alpha e^{-\beta x}, \quad p_\eta(x) = c_2 x^\gamma e^{-\beta x} & \text{for } x > 0 \end{aligned}$$

Find:

- (a) the constants c_1 and c_2 ;
- (b) the density function of the sum $\xi + \eta$.

15. Find the distribution function of the sum of the independent random variables ξ and η , the first of which is uniformly distributed in the interval $(-h, h)$, and the second has the distribution function $F(x)$.

16. The density function of the random vector (ξ, η, ζ) is

$$p(x, y, z) = \begin{cases} \frac{6}{(1+x+y+z)^4} & \text{for } x > 0, y > 0, z > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the distribution of the variable $\xi + \eta + \zeta$.

17. Find the distribution of the sum of the independent random variables ξ_1 and ξ_2 if their distributions are given by the conditions:

(a) $F_1(x) = F_2(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$;

(b) uniform distribution in the intervals $(-5, 1)$, $(1, 5)$, respectively;

(c) $p_1(x) = p_2(x) = \frac{1}{2\alpha} e^{-\frac{|x|}{\alpha}}$.

18. The density function of the independent random variables ξ and η is:

(a) $p_\xi(x) = p_\eta(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ ae^{-ax} & \text{for } x > 0 \quad (a > 0) \end{cases}$

(b) $p_\xi(x) = p_\eta(x) = \begin{cases} 0 & \text{for } x \leq 0, x > a \\ \frac{1}{a} & \text{for } 0 < x \leq a \end{cases}$

Find the density function of the variable $\zeta = \frac{\xi}{\eta}$.

19. Find the distribution function of the product of the independent factors ξ and η on the basis of their distribution functions $F_1(x)$ and $F_2(x)$.

20. The random variables ξ and η are independent and distributed as follows:

(a) uniformly in the interval $(-a, a)$;

(b) normally with parameters $a=0$, $\sigma=1$.

Find the distribution function of their product.

21. The sides ξ and η of a triangle are independent random variables. Using their distribution functions $F_\xi(x)$ and $F_\eta(x)$ find the distribution function of the third side if the angle between the sides ξ and η is equal to a constant number α .

22. Prove that if the variables ξ and η are independent and their density function is

$$p_\xi(x) = p_\eta(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ e^{-x} & \text{for } x > 0 \end{cases}$$

then the variables $\xi + \eta$ and $\frac{\xi}{\eta}$ are also independent.

23. Prove that if the variables ξ and η are independent and normally distributed with parameters $a_1 = a_2 = 0$, $\sigma_1 = \sigma_2 = \sigma$, then the variables

$$\zeta = \xi^2 + \eta^2 \quad \text{and} \quad \delta = \frac{\xi}{\eta}$$

are also independent.

24. Prove that if the variables ξ and η are independent and distributed in accordance with the chi-square law with parameters m and n , then the variables $\delta = \frac{\xi}{\eta}$ and $\zeta = \xi + \eta$ are independent.

25. The random variables $\xi_1, \xi_2, \dots, \xi_n$ are independent and have one and the same density function

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Find the two-dimensional density function of the variables

$$\eta = \sum_{k=1}^n \xi_k \quad \text{and} \quad \zeta = \sum_{k=1}^m \xi_k \quad (m < n)$$

26. Prove that any distribution function possesses the following properties:

$$\begin{aligned} \lim_{x \rightarrow \infty} x \int_x^{\infty} \frac{1}{z} dF(z) &= 0, & \lim_{x \rightarrow +0} x \int_x^{\infty} \frac{1}{z} dF(z) &= 0, \\ \lim_{x \rightarrow -\infty} x \int_{-\infty}^x \frac{1}{z} dF(z) &= 0, & \lim_{x \rightarrow -0} x \int_{-\infty}^x \frac{1}{z} dF(z) &= 0 \end{aligned}$$

27. Two series of independent trials are performed with a random variable ξ which has a continuous distribution function $F(x)$. As a result, ξ took on values arranged in the order of increasing magnitude in each series:

$$x_1 < x_2 < \dots < x_M, \quad y_1 < y_2 < \dots < y_N$$

What is the probability of the inequalities

$$y_\mu < x_{m+1} < y_{\mu+1}$$

where m and μ are given numbers ($0 < m < M$, $0 < \mu < N$)?

28. The random variable ξ has a continuous distribution function $F(x)$. As a result of n independent observations of ξ , we have the following values $x_1 < x_2 < \dots < x_n$ that are arranged in increasing order of magnitude. Find the density function of the variable

$$\eta = \frac{F(x_n) - F(x_2)}{F(x_n) - F(x_1)}$$

29. The random variables ξ and η are independent and identically distributed with the density function

$$p_\xi(x) = p_\eta(x) = \frac{C}{1+x^4}$$

Find the constant C and prove that the variable $\frac{\xi}{\eta}$ is distributed in accordance with the Cauchy law.

30. The random variables ξ and η are independent and their density functions are, respectively, given by

$$p_\xi(x) = \frac{1}{\pi \sqrt{1-x^2}} \quad (|x| < 1) \quad \text{and} \quad p_\eta(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ xe^{-\frac{x^2}{2}} & \text{for } x > 0 \end{cases}$$

Prove that the variable $\xi\eta$ is normally distributed.

31. Let ξ and ζ be independent and let them have the density functions

$$p_\xi(x) = p_\zeta(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \lambda e^{-\lambda x} & \text{for } x > 0 \end{cases}$$

Prove that the relation $\eta = \frac{\xi}{\xi + \zeta}$ is distributed uniformly on the interval $(0, 1)$.

32. The random variables ξ and η are independent and are uniformly distributed on the interval $(-1, 1)$. Compute the probability that the roots of the equation $x^2 + \xi x + \eta = 0$ are real.

(The problems from 29 to 32 were communicated to me by M. I. Yadrenko.)

CHAPTER 5

Numerical Characteristics of Random Variables

In the preceding chapter we saw that the fullest description of a random variable is given by its distribution function. Indeed, the distribution function indicates at the same time both what values the random variable can assume and with what probabilities. However, in a number of cases we need to know much less about the random variable, we want merely a general idea. Very important in the theory of probability and its applications are certain constant numbers that are obtained in accordance with specific rules from the distribution functions of random variables. Of these constants which serve to give a general quantitative description of random variables, of particular importance are mathematical expectation, variance and moments of various orders.

Sec. 26. Mathematical Expectation

We begin by considering the following schematic example: suppose that when firing from a certain gun, it is necessary to fire one shell with a probability of p_1 to hit the target, two shells with a probability p_2 , three shells with a probability p_3 , and so forth. Also, it is known that n shells are definitely sufficient to hit the target. We thus know that

$$p_1 + p_2 + \dots + p_n = 1$$

Now, how many shells, on the average, are needed to hit the target? We reason as follows. Suppose that a very large number of shots are fired under the conditions stated above. Then on the basis of the Bernoulli theorem we can assert that the relative number of shots in which only one shell would suffice to hit the target is approximately equal to p_1 . In exactly the same way, two shells would require approximate-

ly $100p_2\%$ shots, and so forth. Thus, "on the average", approximately

$$1 \cdot p_1 + 2 \cdot p_2 + \dots + n \cdot p_n$$

shells will be needed to hit one target.

Similar problems involving the computing of the average value of a random variable crop up in a great diversity of problems. That is why a special constant, called *mathematical expectation*, is introduced into probability theory. We shall first give a definition for discrete random variables by proceeding from the foregoing example.

Let

$$x_1, x_2, \dots, x_n, \dots$$

denote possible values of a discrete random variable ξ , and let

$$p_1, p_2, \dots, p_n, \dots$$

denote the corresponding probabilities.

If the series $\sum_{n=1}^{\infty} x_n p_n$ converges *absolutely*, then its sum is called the *mathematical expectation* (or, simply, *expectation*) of the random variable ξ and is denoted by $M\xi$.

For continuous random variables, it will be natural to give the following definition: if a random variable ξ is continuous and $p(x)$ is its density function, then the expectation of ξ is the integral

$$M\xi = \int x p(x) dx \quad (1)$$

in those cases when the integral

$$\int |x| p(x) dx$$

exists.

For an arbitrary random variable ξ with distribution function $F(x)$, the expectation is the integral

$$M\xi = \int x dF(x) \quad (2)$$

Taking advantage of the Stieltjes integral, we can give a simple geometrical interpretation of the notion of expectation: the expectation is equal to the difference between the areas bounded by the y -axis, the straight line $y=1$, and the curve $y=F(x)$ in the interval $(0, +\infty)$ and bounded by the x -axis, the curve $y=F(x)$ and the y -axis in the interval $(-\infty, 0)$. In Fig. 18 the appropriate areas are shaded and the sign is indicated that is to be affixed to the sum of each area. Let us point out, incidentally, that the geometrical illustration per-

mits us to write the expectation in the following form:

$$M\xi = - \int_{-\infty}^0 F(x) dx + \int_0^{\infty} (1 - F(x)) dx \quad (3)$$

This remark makes it possible, in many cases, to find the expectation almost without any computations. For instance, the expectation of the random variable distributed according to the law given at the end of Sec. 22 is one half.

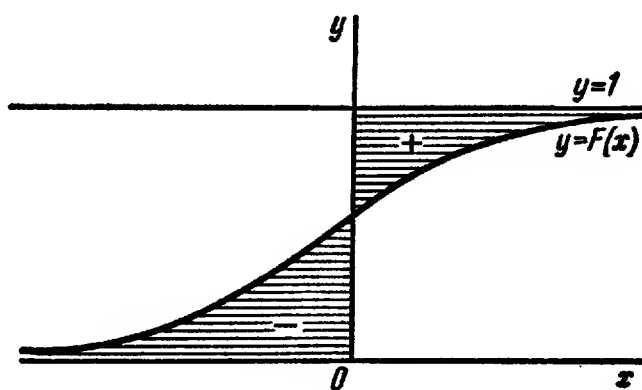


Fig. 18

Note that of the earlier considered random variables, the one distributed in accordance with the Cauchy law (Example 5, Sec. 24) does not have any expectation.

Let us now consider some examples.

Example 1. Find the expectation of the random variable distributed according to the normal law

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$$

From formula (2) we find

$$M\xi = \int x \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2\sigma^2}\right) dx$$

The change of variables $z = \frac{x-a}{\sigma}$ reduces the integral to the form

$$M\xi = \frac{1}{\sqrt{2\pi}} \int (\sigma z + a) e^{-\frac{z^2}{2}} dz = \frac{\sigma}{\sqrt{2\pi}} \int z e^{-\frac{z^2}{2}} dz + \frac{a}{\sqrt{2\pi}} \int e^{-\frac{z^2}{2}} dz$$

Since

$$\int e^{-\frac{z^2}{2}} dz = \sqrt{2\pi} \quad \text{and} \quad \int z e^{-\frac{z^2}{2}} dz = 0$$

it follows that

$$M\xi = a$$

We have obtained an important result that elucidates the probabilistic meaning of one of the parameters defining the normal law: *in the normal law of distribution the parameter a is equal to the expectation.*

Example 2. Determine the expectation of the random variable ξ uniformly distributed in the interval (a, b) .

We have

$$M\xi = \int_a^b x \frac{dx}{b-a} = \frac{b^2 - a^2}{2(b-a)} = \frac{a+b}{2}$$

We see that the expectation coincides with the midpoint of the interval of possible values of the random variable.

Example 3. Determine the expectation of the random variable ξ which is distributed in accordance with the Poisson law

$$P\{\xi = k\} = \frac{a^k e^{-a}}{k!} \quad (k = 0, 1, 2, \dots)$$

We have

$$M\xi = \sum_{k=0}^{\infty} k \cdot \frac{a^k e^{-a}}{k!} = \sum_{k=1}^{\infty} k \cdot \frac{a^k e^{-a}}{k!} = a e^{-a} \sum_{k=1}^{\infty} \frac{a^{k-1}}{(k-1)!} = a e^{-a} \sum_{k=0}^{\infty} \frac{a^k}{k!} = a$$

If $F(x/B)$ is the conditional distribution function for a random variable ξ , then we will call the integral

$$M(\xi/B) = \int x dF(x/B) \quad (4)$$

the conditional expectation of the random variable ξ with respect to the event B .

Let B_1, B_2, \dots, B_n be a complete group of mutually exclusive events and $F(x/B_1), F(x/B_2), \dots, F(x/B_n)$ the conditional distribution functions of the variable ξ corresponding to these events. Let $F(x)$ denote the unconditional distribution function of ξ ; using the formula of total probability, we find

$$F(x) = \sum_{k=1}^n P(B_k) F(x/B_k)$$

Together with (4), this equation enables us to obtain the following formula:

$$M\xi = \sum_{k=1}^n P(B_k) M(\xi/B_k)$$

which, obviously, may be written differently:

$$M\xi = M\{M(\xi/B_k)\} \quad (5)$$

The foregoing formula greatly simplifies in many cases the calculation of expectations.

Example 4. A workman is operating n machines of one type arranged in a straight line at separations a from one another (Fig. 19). Assuming that the operator moves from machine to machine in order of priority, find the average path length (the expectation of the path length) between machines.

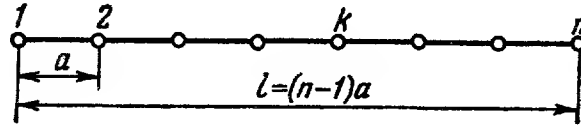


Fig. 19

Number the machines from left to right, 1 to n , and denote by B_k the event that the operator is at the k th machine. Since all the machines are of the same kind, the probability $p_i^{(k)}$ that the next machine requiring the attention of the operator will be the i th is equal to $\frac{1}{n}$ ($1 \leq i \leq n$). The path length λ in this case is

$$\lambda_i^{(k)} = \begin{cases} (k-i)a & \text{for } k \geq i \\ (i-k)a & \text{for } k < i \end{cases}$$

By definition

$$\begin{aligned} M(\lambda/B_k) &= \frac{1}{n} \left(\sum_{i=1}^k (k-i)a + \sum_{i=k+1}^n (i-k)a \right) = \\ &= \frac{a}{n} \left(\frac{k(k-1)}{2} + \frac{(n-k)(n-k+1)}{2} \right) = \\ &= \frac{a}{2n} [2k^2 - 2(n+1)k + n(n+1)] \end{aligned}$$

The probability that the operator will be at k th machine is $1/n$, and so from formula (5) we find

$$M\lambda = \sum_{k=1}^n \frac{a}{2n^2} [2k^2 - 2(n+1)k + n(n+1)]$$

We know that

$$\sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}$$

and so

$$M\lambda = \frac{a(n^2-1)}{3n} = \frac{l}{3} \left(1 + \frac{1}{n} \right)$$

where $l = (n-1)a$ denotes the distance between the end machines.

The expectation of an n -dimensional random variable $(\xi_1, \xi_2, \dots, \xi_n)$ is defined as the collection of n integrals:

$$a_k = \int \int \dots \int x_k dF(x_1, \dots, x_k, \dots, x_n) = \int x dF_k(x) = M\xi_k$$

where $F_k(x)$ is the distribution function of the variable ξ_k .*

Example 5. The density function of a two-dimensional random variable (ξ_1, ξ_2) is given by the formula (two-dimensional normal distribution)

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x_1-a)^2}{\sigma_1^2} - \frac{2r(x_1-a)(x_2-b)}{\sigma_1\sigma_2} + \frac{(x_2-b)^2}{\sigma_2^2} \right] \right\}$$

Find its expectation.

By definition,

$$a_1 = \int \int x_1 p(x_1, x_2) dx_2 = \int x_1 p_1(x_1) dx_1$$

and

$$a_2 = \int \int x_2 p(x_1, x_2) dx_1 dx_2 = \int x_2 p_2(x_2) dx_2$$

In Example 2 of Sec. 23 we saw that

$$p_1(x_1) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{(x_1-a)^2}{2\sigma_1^2} \right\}$$

$$p_2(x_2) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp \left\{ -\frac{(x_2-b)^2}{2\sigma_2^2} \right\}$$

and so from the results of Example 1 of this section we find

$$a_1 = a \text{ and } a_2 = b$$

We have also been able to find the probabilistic meaning of parameters a and b for a two-dimensional normal distribution.

Sec. 27. Variance

The *variance* (also called *dispersion*) of a random variable ξ is defined as the expectation of the square of the deviation of ξ from $M\xi$. Let us agree to denote the variance by the symbol $D\xi$.

* We do not give a formal definition of an n -dimensional Stieltjes integral, firstly, because we will actually only consider discrete and continuous random variables and, secondly, because probability theory does not require a general theory of Stieltjes integrals but the theory of the abstract Lebesgue integral (for more details see Chapter I of the monograph *Limit Distributions for Sums of Independent Random Variables* by Gnedenko and Kolmogorov, 1949, in Russian).

Then by definition

$$D\xi = M(\xi - M\xi)^2 = \int_0^{\infty} x dF_{\eta}(x) \quad (1)$$

where $F_{\eta}(x)$ denotes the distribution function of the random variable $\eta = (\xi - M\xi)^2$.

For practical calculations, use is made of a different formula, namely

$$D\xi = \int (z - M\xi)^2 dF_{\xi}(z) \quad (2)$$

That formulas (1) and (2) are equivalent follows directly from the following proposition.

Theorem. *If $F_{\xi}(x)$ is the distribution function of a variable ξ and $f(x)$ is a continuous function, then*

$$Mf(\xi) = \int f(x) dF_{\xi}(x)$$

We shall confine the proof of this theorem only to the most elementary special case: $f(x) = (x - a)^k$. Using the notation

$$G(x) = P\{(\xi - a)^k < x\}$$

we find by definition that

$$M(\xi - a)^k = \int_{-\infty}^{\infty} x dG(x)$$

If k is an odd number, then $(\xi - a)^k$ is a nondecreasing function of ξ and therefore

$$\begin{aligned} G(x) = P\{(\xi - a)^k < x\} &= P\{\xi - a < \sqrt[k]{x}\} = \\ &= P\{\xi < a + \sqrt[k]{x}\} = F(a + \sqrt[k]{x}) \end{aligned}$$

Thus, for odd k ,

$$M(\xi - a)^k = \int x dF(a + \sqrt[k]{x})$$

It is easy to see that by substituting $z = a + \sqrt[k]{x}$ we reduce this integral to the form

$$M(\xi - a)^k = \int_{-\infty}^{\infty} (x - a)^k dF(x)$$

But if k is even, then $(\xi - a)^k$ is a nonnegative quantity and, hence, $G(x) = 0$ for $x \leq 0$. For $x > 0$

$$\begin{aligned} G(x) = P\{(\xi - a)^k < x\} &= P\{a - \sqrt[k]{x} < \xi < a + \sqrt[k]{x}\} = \\ &= F(a + \sqrt[k]{x}) - F(a - \sqrt[k]{x} + 0) \end{aligned}$$

Thus, for even k

$$M(\xi - a)^k = \int_0^\infty x dF(a + \sqrt[k]{x}) - \int_0^\infty x dF(a - \sqrt[k]{x} + 0)$$

By the substitutions $z = a + \sqrt[k]{x}$ in the first integral and $z = a - \sqrt[k]{x}$ in the second one, we reduce $M(\xi - a)^k$ to the form

$$M(\xi - a)^k = \int_{-\infty}^\infty (x - a)^k dF(x)$$

For practical purposes, it is useful to write formula (2) in a different form. Since

$$(z - M\xi)^2 = z^2 - 2zM\xi + (M\xi)^2 \quad \text{and} \quad M\xi = \int z dF_\xi(z)$$

it follows that formula (2) may be written differently:

$$D\xi = \int z^2 dF_\xi(z) - \left(\int z dF_\xi(z) \right)^2 = M\xi^2 - (M\xi)^2 \quad (3)$$

Since the variance is a nonnegative quantity, from this relation we derive

$$\int z^2 dF_\xi(z) \geq \left(\int z dF_\xi(z) \right)^2$$

This inequality is a particular case of the well-known Bunyakovsky-Cauchy inequality (also called the Schwarz inequality).

Like expectation, variance does not exist for all random variables. For instance, the Cauchy law which we considered earlier (see Example 5, Sec. 24) does not have a finite variance.

Let us consider some examples in computing variance.

Example 1. Find the variance of a random variable ξ uniformly distributed in the interval (a, b) .

In our example,

$$\int x^2 dF_\xi(x) = \int_a^b \frac{x^2}{b-a} dx = \frac{b^3 - a^3}{3(b-a)} = \frac{b^2 + ab + a^2}{3}$$

In the preceding section,

$$M\xi = \frac{a+b}{2}$$

was found.

Thus,

$$D\xi = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2} \right)^2 = \frac{(b-a)^2}{12}$$

We see that the variance depends only on the length of the interval (a, b) and is an increasing function of the length. The greater the interval of values which the random variable assumes, i.e., the more the values are scattered, the greater the variance. Variance is thus a *measure of the spread* or *dispersion* of the values of a random variable about the expectation.

Example 2. Find the variance of the random variable ξ distributed in accord with the normal law

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(x-a)^2}{2\sigma^2} \right\}$$

We know that $M\xi = a$, and therefore

$$D\xi = \int (x-a)^2 p(x) dx = \frac{1}{\sigma \sqrt{2\pi}} \int (x-a)^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx$$

Changing variables in the integral, put

$$z = \frac{x-a}{\sigma}$$

then

$$D\xi = \frac{\sigma^2}{\sqrt{2\pi}} \int z^2 e^{-\frac{z^2}{2}} dz$$

Integrating by parts we find

$$\int z^2 e^{-\frac{z^2}{2}} dz = \int_{-\infty}^{\infty} -ze^{-\frac{z^2}{2}} + \int e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

And so, finally,

$$D\xi = \sigma^2$$

We have thus found the probabilistic meaning of the second parameter that determines the normal law. We see that the *normal law of distribution is fully determined by the expectation and the variance*. This is widely used in theoretical investigations.

It will be noted that in the case of a normally distributed random variable the variance permits one to judge about the dispersion of its values. Though for any positive values of variance, a normally distributed random variable can assume all real values, still, the dispersion of values of the variable will be the less, the smaller the variance. Here, the probabilities of values close to the expectation will be greater. We noted this fact in the preceding chapter when we first examined the normal law.

Example 3. Find the variance of the random variable λ considered in Example 4, Sec. 26.

Retaining the notations of Example 4, we find

$$\begin{aligned} \mathbf{M}(\lambda^2/B_k) &= \frac{1}{n} \left(\sum_{i=1}^k (k-i)^2 a^2 + \sum_{i=k+1}^n (i-k)^2 a^2 \right) = \\ &= \frac{a^2}{6n} [(k-1) \cdot k (2k-1) + (n-k)(n-k+1)(2n-2k+1)] = \\ &= \frac{a^2}{6} [6k^2 - 6(n+1)k + (2n+1)(n+1)] \end{aligned}$$

and, consequently,

$$\begin{aligned} \mathbf{M}(\lambda^2) &= \frac{1}{n} \sum_{k=1}^n \mathbf{M}(\lambda^2/B_k) = \\ &= \frac{a^2}{6n} [n(n+1)(2n+1) - 3(n+1)^2 n + n(n+1)(2n+1)] = \frac{a^2}{6} (n^2 - 1) \end{aligned}$$

From this it follows that

$$\begin{aligned} \mathbf{D}(\lambda) &= \mathbf{M}(\lambda^2) - \mathbf{M}(\lambda)^2 = \frac{a^2}{6} (n^2 - 1) - \frac{a^2 (n^2 - 1)^2}{9n^2} = \\ &= \frac{a^2 (n^2 - 1)(n^2 + 2)}{18n^2} = \frac{l^2}{18} \left(1 + \frac{2}{n} + \frac{4}{n^2} + \frac{6}{n^2(n-1)} \right) \end{aligned}$$

The *variance* (or the *covariance matrix*) of an n -dimensional random variable $(\xi_1, \xi_2, \dots, \xi_n)$ is defined as the set of n^2 constants given by the formula

$$\begin{aligned} b_{jk} &= \int \int \dots \int (x_j - \mathbf{M}\xi_j)(x_k - \mathbf{M}\xi_k) dF(x_1, x_2, \dots, x_n) \quad (4) \\ &\quad (1 \leq k \leq n, \quad 1 \leq j \leq n) \end{aligned}$$

Since for any real t_j ($1 \leq j \leq n$)

$$\int \dots \int \left\{ \sum_{j=1}^n t_j (x_j - \mathbf{M}\xi_j) \right\}^2 dF(x_1, x_2, \dots, x_n) = \sum_{j=1}^n \sum_{k=1}^n b_{jk} t_j t_k \geq 0$$

it follows that, as we know from the theory of quadratic forms, the quantities b_{jk} satisfy the inequalities

$$\begin{vmatrix} b_{11} & b_{12} & \dots & b_{1k} \\ b_{21} & b_{22} & \dots & b_{2k} \\ \dots & \dots & \dots & \dots \\ b_{k1} & b_{k2} & \dots & b_{kk} \end{vmatrix} \geq 0 \quad \text{for } k=1, 2, \dots, n$$

It is obvious that

$$b_{kk} = \mathbf{D}\xi_k$$

The quantities b_{jk} for $k \neq j$ are called the *mixed central moments of the second order* of the variables ξ_j and ξ_k ; obviously,

$b_{jk} = b_{kj}$. In the statistical literature, b_{jk} is often called the *covariance* of ξ_j and ξ_k and is denoted by the symbol $\text{cov}(\xi_j, \xi_k)$.

The following function of second-order moments

$$r_{ij} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}}$$

is called the *correlation coefficient* of the variables ξ_i and ξ_j .

The magnitude of the correlation coefficient lies within the limits $(-1, +1)$.

The correlation coefficient r_{ij} assumes the values ± 1 only when ξ_j and ξ_k are connected by a linear relationship.

Indeed, since

$$D\left(\frac{\xi_i}{\sqrt{b_{ii}}} \pm \frac{\xi_j}{\sqrt{b_{jj}}}\right) = 2(1 \pm r_{ij}) \geq 0$$

it follows that $-1 \leq r_{ij} \leq 1$.

The equality $r_{ij} = 1$ is possible if and only if

$$D\left(\frac{\xi_i}{\sqrt{b_{ii}}} - \frac{\xi_j}{\sqrt{b_{jj}}}\right) = 0$$

Now the variance can be zero only for random variables which assume a certain constant value with probability one. Thus, if $r_{ij} = 1$, then

$$\frac{\xi_i}{\sqrt{b_{ii}}} - \frac{\xi_j}{\sqrt{b_{jj}}} = c$$

and, hence,

$$\xi_i = \sqrt{\frac{b_{ii}}{b_{jj}}} \xi_j + \alpha \quad (\alpha = c\sqrt{b_{ii}})$$

In exactly the same way, if $r_{ij} = -1$, then

$$\xi_i = -\sqrt{\frac{b_{ii}}{b_{jj}}} \xi_j + \alpha$$

By straightforward computation it is proved that *if random variables are linearly related, their correlation coefficient is equal to plus or minus unity*.

It is easy to compute that *for independent random variables ξ_i and ξ_j the correlation coefficient is zero*.

The converse conclusion is not true. The correlation coefficient of the variables ξ and η may be zero even though they are dependent. To illustrate, suppose $\eta = \xi^2$, ξ is distributed symmetrically about the point $x=0$ and has a finite fourth moment. Then $M\xi=0$, $M\xi\eta = M\xi^3=0$, consequently, $M(\xi - M\xi)(\eta - M\eta) = 0$ and, hence, $r_{\xi\eta} = 0$. We can thus say that the correlation coefficient is a measure of the

strength of the relationship (linear relation) between the variables ξ_i and ξ_j .

Example 4. Find the variance of the two-dimensional random variable (ξ_1, ξ_2) distributed in accordance with the nondegenerate normal law

$$p(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \\ \times \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right] \right\}$$

According to formula (4) and the results of Example 2 of this section and Example 1 of Sec. 26 we find

$$D\xi_1 = \sigma_1^2, \quad D\xi_2 = \sigma_2^2$$

Further,

$$b_{12} = b_{21} = \iint (x-a)(y-b) p(x, y) dx dy = \\ = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \int e^{-\frac{(y-b)^2}{2\sigma_2^2}} dy \times \\ \times \int (x-a)(y-b) \exp \left\{ -\frac{1}{2(1-r^2)} \left(\frac{x-a}{\sigma_1} - r \frac{y-b}{\sigma_2} \right)^2 \right\} dx$$

By the substitutions $z = \frac{1}{\sqrt{1-r^2}} \left(\frac{x-a}{\sigma_1} - r \frac{y-b}{\sigma_2} \right)$, $t = \frac{y-b}{\sigma_2}$ the expression for b_{12} is reduced to the form

$$b_{12} = b_{21} = \frac{1}{2\pi} \iint (\sigma_1\sigma_2\sqrt{1-r^2}tz + r\sigma_1\sigma_2t^2) e^{-\frac{t^2}{2}-\frac{z^2}{2}} dz dt = \\ = \frac{r\sigma_1\sigma_2}{2\pi} \int t^2 e^{-\frac{t^2}{2}} dt \int e^{-\frac{z^2}{2}} dz + \\ + \frac{\sigma_1\sigma_2\sqrt{1-r^2}}{2\pi} \int t e^{-\frac{t^2}{2}} dt \int z e^{-\frac{z^2}{2}} dz = r\sigma_1\sigma_2$$

Whence we find

$$r = \frac{\iint (x-a)(y-b) p(x, y) dx dy}{\sigma_1\sigma_2} = \frac{M(\xi_1 - M\xi_1)(\xi_2 - M\xi_2)}{\sqrt{D\xi_1 D\xi_2}}$$

Summarizing, then, the parameter r of a two-dimensional normal distribution is the correlation coefficient of the components (ξ_1, ξ_2) .

We see that the *two-dimensional normal law*, like in the one-dimensional case, is completely determined by specifying the expectation and the variance; that is, it is determined by specifying five quantities: $M\xi_1$, $M\xi_2$, $D\xi_1$, $D\xi_2$ and r .

Sec. 28. Theorems on Expectation and Variance

Theorem 1. *The expectation of a constant is equal to that constant.*

Proof. We can regard the constant C as a discrete random variable which can take on only one value C with probability one; and so

$$MC = C \cdot 1 = C$$

Theorem 2. *The expectation of a sum of random variables is equal to the sum of their expectations:*

$$M(\xi + \eta) = M\xi + M\eta$$

Proof. First consider the case of the discrete random variables ξ and η . Let $a_1, a_2, \dots, a_n, \dots$ be possible values of the variable ξ and $p_1, p_2, \dots, p_n, \dots$ be the probabilities of these values; $b_1, b_2, \dots, b_k, \dots$ are possible values of the variable η and $q_1, q_2, \dots, q_k, \dots$ are the probabilities of these values. The possible values of the variable $\xi + \eta$ have the form $a_n + b_k$ ($k, n = 1, 2, \dots$). Denote by p_{nk} the probability that ξ will assume the value a_n and by η , the value b_k . By the definition of expectation,

$$\begin{aligned} M(\xi + \eta) &= \sum_{n, k=1}^{\infty} (a_n + b_k) p_{nk} = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} (a_n + b_k) p_{nk} = \\ &= \sum_{n=1}^{\infty} a_n \left(\sum_{k=1}^{\infty} p_{nk} \right) + \sum_{k=1}^{\infty} b_k \left(\sum_{n=1}^{\infty} p_{nk} \right) \end{aligned}$$

Since by the theorem of total probability

$$\sum_{k=1}^{\infty} p_{nk} = p_n \quad \text{and} \quad \sum_{n=1}^{\infty} p_{nk} = q_k$$

it follows that

$$\sum_{n=1}^{\infty} a_n \sum_{k=1}^{\infty} p_{nk} = \sum_{n=1}^{\infty} a_n p_n = M\xi$$

and

$$\sum_{k=1}^{\infty} b_k \sum_{n=1}^{\infty} p_{nk} = \sum_{k=1}^{\infty} p_k q_k = M\eta$$

The proof of the theorem for the case of discrete summands is complete.

The same applies to the case when there is a two-dimensional density function $p(x, y)$ of a random variable (ξ, η) ; from formula (3)

of Sec. 24 we find

$$\begin{aligned} M\xi &= M(\xi + \eta) = \int x dF_{\xi}(x) = \int x \left(\int p(z, x-z) dz \right) dx = \\ &= \int \int xp(z, x-z) dz dx = \int \int (z+y) p(z, y) dz dy = \\ &= \int \int zp(z, y) dz dy + \int \int yp(z, y) dz dy = \\ &= \int zp_{\xi}(z) dz + \int yp_{\eta}(y) dy = M\xi + M\eta \end{aligned}$$

Theorem 2 will be proved for the general case in Sec. 29.

Corollary 1. *The expectation of the sum of a finite number of random variables is equal to the sum of their expectations:*

$$M(\xi_1 + \xi_2 + \dots + \xi_n) = M\xi_1 + M\xi_2 + \dots + M\xi_n$$

Indeed, by virtue of the theorem we have just proved

$$\begin{aligned} M(\xi_1 + \xi_2 + \dots + \xi_n) &= M\xi_1 + M(\xi_2 + \xi_3 + \dots + \xi_n) = \\ &= M\xi_1 + M\xi_2 + M(\xi_2 + \dots + \xi_n) = \dots = M\xi_1 + M\xi_2 + \dots + M\xi_n \end{aligned}$$

Corollary 2. *Consider the sum*

$$\zeta_{\mu} = \xi_1 + \xi_2 + \dots + \xi_{\mu}$$

where μ is a random variable that takes on only integral values, the random variables ξ_1, ξ_2, \dots do not depend on μ , the expectation of μ is finite and the series

$$\sum_{k=1}^{\infty} M|\xi_k| P\{\mu \geq k\}$$

converges; the expectation of the sum exists and is equal to

$$M\zeta_{\mu} = \sum_{j=1}^{\infty} M\xi_j P\{\mu \geq j\}$$

Proof. Indeed, provided that $\mu = k$, the conditional expectation is

$$M\{\zeta_{\mu}/\mu = k\} = M\xi_1 + M\xi_2 + \dots + M\xi_k$$

The unconditional expectation is

$$\begin{aligned} M\zeta_{\mu} &= \sum_{k=1}^{\infty} M\{\zeta_{\mu}/\mu = k\} \cdot P(\mu = k) = \sum_{k=1}^{\infty} P\{\mu = k\} \sum_{j=1}^k M\xi_j = \\ &= \sum_{j=1}^{\infty} M\xi_j \sum_{k=j}^{\infty} P\{\mu = k\} = \sum_{j=1}^{\infty} M\xi_j P\{\mu \geq j\} \end{aligned}$$

If the summands $\xi_1, \xi_2, \xi_3, \dots$ are identically distributed, that is, if $P\{\xi_1 < x\} = P\{\xi_2 < x\} = \dots = F(x)$, then

$$M_{\zeta_\mu} = M_{\xi_1} \cdot M_\mu$$

Indeed,

$$M_{\zeta_\mu} = \sum_{k=1}^{\infty} P\{\mu = k\} \sum_{j=1}^k M_{\xi_j} = M_{\xi_1} \sum_{k=1}^{\infty} k P\{\mu = k\} = M_{\xi_1} \cdot M_\mu$$

Example 1. The number of cosmic particles striking a given area is a random variable μ that obeys the Poisson law with parameter a ; each of the particles carries energy ξ that depends on chance. Find the mean energy \mathcal{E} acquired by the area in unit time.

According to Corollary 2 we have

$$M_{\mathcal{E}} = M_{\xi} \cdot M_\mu = a M_{\xi}$$

Example 2. A target is fired at and hit n times. Assuming that the shots are fired independently of one another and the probability of a hit in each shot is p , find the expectation of shell consumption.

Denote by ξ_k the number of shells expended from the $(k-1)$ st hit to the k th hit. It is obvious that the consumption of shells in n hits is

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

and, consequently,

$$M_{\xi} = M_{\xi_1} + M_{\xi_2} + \dots + M_{\xi_n}$$

But

$$M_{\xi_1} = M_{\xi_2} = \dots = M_{\xi_n}$$

and

$$M_{\xi_1} = \sum_{k=0}^{\infty} k q^{k-1} p = \frac{p}{(1-q)^2} = \frac{1}{p}$$

consequently,

$$M_{\xi} = \frac{n}{p}$$

Theorem 3. The expectation of the product of the independent random variables ξ and η is equal to the product of their expectations.

Proof. If the variables ξ and η are discrete, $a_1, a_2, \dots, a_k, \dots$ are the possible values of ξ and $p_1, p_2, \dots, p_k, \dots$ are the probabilities of these values, $b_1, b_2, \dots, b_n, \dots$ are the possible values of η and $q_1, q_2, \dots, q_n, \dots$ are the probabilities of these values, then the probability that ξ will assume the value a_k and η will assume the value

b_n is $p_k q_n$. By the definition of expectation

$$\begin{aligned} M\xi\eta &= \sum_{k,n} a_k b_n p_k q_n = \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} a_k b_n p_k q_n = \\ &= \left(\sum_{k=1}^{\infty} a_k p_k \right) \left(\sum_{n=1}^{\infty} b_n q_n \right) = M\xi M\eta \end{aligned}$$

Only slightly more complicated is the proof for the case of continuous variables. This is left to the reader.

Proof of the theorem in the general case will be taken up in Sec. 29.

Corollary 1. *A constant factor may be removed from under the sign of expectation:*

$$MC\xi = CM\xi$$

This assertion is obvious, since no matter what the variable ξ , the constant C and the variable ξ may be regarded as independent variables.

Theorem 4. *The variance of a constant is zero.*

Proof. According to Theorem 1,

$$DC = M(C - MC)^2 = M(C - C)^2 = M0 = 0$$

Theorem 5. *If C is constant, then*

$$DC\xi = C^2 D\xi$$

Proof. By virtue of the corollary to Theorem 3,

$$\begin{aligned} DC\xi &= M[C\xi - MC\xi]^2 = M[C\xi - CM\xi]^2 = \\ &= MC^2 [\xi - M\xi]^2 = C^2 M[\xi - M\xi]^2 = C^2 D\xi \end{aligned}$$

Theorem 6. *The variance of the sum of the independent random variables ξ and η is equal to the sum of their variances:*

$$D(\xi + \eta) = D\xi + D\eta$$

Proof. Indeed,

$$\begin{aligned} D(\xi + \eta) &= M[\xi + \eta - M(\xi + \eta)]^2 = M[(\xi - M\xi) + (\eta - M\eta)]^2 = \\ &= D\xi + D\eta + 2M(\xi - M\xi)(\eta - M\eta) \end{aligned}$$

The variables ξ and η are independent, and so also independent are the quantities $\xi - M\xi$ and $\eta - M\eta$; whence

$$M(\xi - M\xi)(\eta - M\eta) = M(\xi - M\xi) M(\eta - M\eta) = 0$$

Corollary 1. *If $\xi_1, \xi_2, \dots, \xi_n$ are random variables, each of which is independent of the sum of the preceding ones, then*

$$D(\xi_1 + \xi_2 + \dots + \xi_n) = D\xi_1 + D\xi_2 + \dots + D\xi_n$$

Corollary 2. *The variance of the sum of a finite number of pair-wise independent random variables $\xi_1, \xi_2, \dots, \xi_n$ is equal to the sum of their variances.*

Proof. Indeed,

$$\begin{aligned} D(\xi_1 + \xi_2 + \dots + \xi_n) &= M \left(\sum_{k=1}^n (\xi_k - M\xi_k) \right)^2 = \\ &= M \sum_{j=1}^n \sum_{k=1}^n (\xi_k - M\xi_k) (\xi_j - M\xi_j) = \sum_{k=1}^n \sum_{j=1}^n M(\xi_k - M\xi_k) (\xi_j - M\xi_j) = \\ &= \sum_{k=1}^n D\xi_k + \sum_{k \neq j} M(\xi_k - M\xi_k) (\xi_j - M\xi_j) \end{aligned}$$

From the independence of any pair of variables ξ_k and ξ_j ($k \neq j$) it follows that for $k \neq j$

$$M(\xi_k - M\xi_k) (\xi_j - M\xi_j) = 0$$

This quite obviously completes the proof.

Example 3. The ratio

$$\frac{\xi - M\xi}{\sqrt{D\xi}}$$

is called the *standard deviation* of a random variable. Prove that

$$D \left(\frac{\xi - M\xi}{\sqrt{D\xi}} \right) = 1$$

Indeed, ξ and $M\xi$, considered as random variables, are independent and for this reason, by virtue of Theorems 5 and 6,

$$D \left(\frac{\xi - M\xi}{\sqrt{D\xi}} \right) = \frac{D\xi + D(-M\xi)}{D\xi} = \frac{D\xi}{D\xi} = 1$$

Example 4. If ξ and η are independent random variables, then

$$D(\xi - \eta) = D\xi + D\eta$$

Indeed, by virtue of Theorems 6 and 7, $D(-\eta) = (-1)^2 D\eta = D\eta$ and $D(\xi - \eta) = D\xi + D\eta$.

Example 5. Theorems 2 and 6 permit of an extremely simple computation of the expectation and variance of the number μ of occurrences of an event A in n independent trials.

Let p_k be the probability of an occurrence of the event A in the k th trial.

Denote by μ_k the number of occurrences of the event A in the k th trial. It is obvious that μ_k is a random variable that takes on values 0 and 1 with probabilities $q_k = 1 - p_k$ and p_k , respectively.

Thus, the variable μ may be represented in the form of a sum:

$$\mu = \mu_1 + \mu_2 + \dots + \mu_n$$

Since

$$\begin{aligned} M\mu_k &= 0 \cdot q_k + 1 \cdot p_k = p_k \\ D\mu_k &= M\mu_k^2 - (M\mu_k)^2 = 0 \cdot q_k + 1 \cdot p_k - p_k^2 = p_k(1 - p_k) = p_k q_k \end{aligned}$$

the proved theorems permit concluding that

$$M\mu = p_1 + p_2 + \dots + p_n$$

and

$$D\mu = p_1 q_1 + p_2 q_2 + \dots + p_n q_n$$

For the case of the Bernoulli scheme, $p_k = p$ and, hence,

$$M\mu = np \quad \text{and} \quad D\mu = npq$$

We then note that this gives

$$M \frac{\mu}{n} = p, \quad D \frac{\mu}{n} = \frac{pq}{n}$$

Example 6. Let us find the expectation and the variance of the number of occurrences of an event E in trials connected in a homogeneous Markov chain.

As before, we denote the number of occurrences of the event E in the k th trial by μ_k . The number of occurrences of the event in n trials is equal to the sum

$$\mu = \mu_1 + \mu_2 + \dots + \mu_n$$

But

$$M\mu = \sum_{k=1}^n M\mu_k = \sum_{k=1}^n p_k$$

According to formula (1') of Sec. 20,

$$p_k = p + (p_1 - p) \delta^{k-1}$$

Thus,

$$M\mu = np + \sum_{k=1}^n (p_1 - p) \delta^{k-1} = np + (p_1 - p) \frac{1 - \delta^n}{1 - \delta}$$

By definition,

$$\begin{aligned} D\mu &= M(\mu - M\mu)^2 = M \left[\sum_{k=1}^n (\mu_k - p_k) \right]^2 = \\ &= \sum_{k=1}^n M(\mu_k - p_k)^2 + 2 \sum_{j>i}^n M(\mu_i - p_i)(\mu_j - p_j) \end{aligned}$$

But

$$D\mu_k = p_k q_k = pq + (q-p)(p_1-p)\delta^{k-1} - (p_1-p)^2 \delta^{2k-2}$$

where

$$q_k = 1 - p_k = q - (p_1 - p)\delta^{k-1}$$

Further,

$$M(\mu_i - p_i)(\mu_j - p_j) = M\mu_i \mu_j - p_i p_j$$

Since the probability of the equality $\mu_i \mu_j = 1$ is obviously $p_i p_j^{(i)}$, it follows that

$$M(\mu_i - p_i)(\mu_j - p_j) = p_i(p_j^{(i)} - p_j)$$

Taking advantage of formulas (1') and (2') of Sec. 20, we find

$$M(\mu_i - p_i)(\mu_j - p_j) = pq\delta^{j-i} + (p_1 - p)(q - p)\delta^{j-1} - (p_1 - p)^2 \delta^{i+j-2}$$

Now

$$\sum_{k=1}^n D\mu_k = npq + (q-p)(p_1-p)\frac{1-\delta^n}{1-\delta} - (p_1-p)^2 \frac{1-\delta^{2n}}{1-\delta^2}$$

and

$$\begin{aligned} \sum_{j>i} M(\mu_i - p_i)(\mu_j - p_j) &= npq \frac{\delta}{1-\delta} - pq \frac{\delta}{1-\delta} \left(1 + \frac{1-\delta^n}{1-\delta}\right) + \\ &+ \frac{(p_1-p)(q-p)}{1-\delta} \left(\frac{\delta-\delta^{n+1}}{1-\delta} - n\delta^n\right) - \frac{(p_1-p)^2 \delta (1-\delta^{n-1})(1-\delta^n)}{(1-\delta)(1-\delta^2)} \end{aligned}$$

Thus,

$$D\mu = npq \frac{1+\delta}{1-\delta} + a_n$$

where a_n is a certain quantity that remains bounded as n increases.

Sec. 29. Mathematical Expectation Defined in the Axiomatics of Kolmogorov

This section should be omitted in a first reading, as it requires extended knowledge in the theory of integration. The general conception presented here is a natural development of the construction of concepts of a random event, of probability, and of a random variable as given by A. N. Kolmogorov (see Secs. 9 and 21). In this interpretation, the concept of expectation leads naturally to the abstract Lebesgue integral.

By definition, the expectation of a random variable $\xi = f(e)$ is the integral

$$M\xi = \int_U f(e) P(de)$$

Under the hypothesis B , the conditional expectation is

$$\mathbf{M}(\xi | B) = \int_U f(e) \mathbf{P}(de | B)$$

It can readily be proved that this definition is equivalent to the following

$$\mathbf{M}(\xi | B) = \int_B f(e) \mathbf{P}(de) \frac{1}{\mathbf{P}(B)}$$

which is often better suited to practical employment.

Let it be remarked that if an event B is representable as the sum of a finite or countable set of disjoint events B_k :

$$B = B_1 + B_2 + \dots$$

then

$$\int_B f(e) \mathbf{P}(de) = \sum_k \int_{B_k} f(e) \mathbf{P}(de)$$

It is useful to note that whereas previously the proof of this theorem on the expectation of a sum required rather lengthy reasoning, now the theorem is a consequence of the formula

$$\int (f + g) \mathbf{P}(de) = \int f \mathbf{P}(de) + \int g \mathbf{P}(de)$$

For the independent random variables ξ and η we previously proved the formula

$$\mathbf{M}(\xi \cdot \eta) = \mathbf{M}\xi \cdot \mathbf{M}\eta \quad (1)$$

only in the case of discrete random variables and in the case of continuous random variables.

In the general case, let us define the discrete random variables ξ_n and η_n by the formulas

$$\begin{aligned} \xi_n &= \frac{m}{n} \quad \text{for} \quad \frac{m}{n} \leq \xi < \frac{m+1}{n} \\ \eta_n &= \frac{k}{n} \quad \text{for} \quad \frac{k}{n} \leq \eta < \frac{k+1}{n} \end{aligned}$$

Then

$$\mathbf{M}(\xi_n \cdot \eta_n) = \mathbf{M}\xi_n \cdot \mathbf{M}\eta_n$$

From well-known theorems on passing to the limit under the sign of the Lebesgue integral we can readily derive that

$$\lim \mathbf{M}\xi_n = \mathbf{M}\xi, \quad \lim \mathbf{M}\eta_n = \mathbf{M}\eta, \quad \lim \mathbf{M}(\xi_n \cdot \eta_n) = \mathbf{M}(\xi \cdot \eta)$$

Thus, formula (1) is proved in the general case.

We shall use the results obtained to derive a formula that generalizes the result of Sec. 28 (Corollary 2). This formula will be obtained from the following theorem proved by A. N. Kolmogorov and Yu. V. Prokhorov.

Given a sequence of random variables

$$\xi_1, \xi_2, \dots, \xi_n, \dots$$

let

$$\zeta_v = \xi_1 + \xi_2 + \dots + \xi_v$$

denote the sum of the first v variables, the number of the summands v itself being a random variable.

Denote by S_m the event that $v = m$ and put

$$p_m = P\{S_m\}, \quad P_n = P\{v \geq n\} = \sum_{m=n}^{\infty} p_m$$

Theorem. *If for $n > m$ the random variable ξ_n and the event S_m are independent, there exist the expectations*

$$a_n = M\xi_n$$

(and, hence, the quantities $c_n = M|\xi_n|$ are finite), and the series

$$\sum_{n=1}^{\infty} c_n P_n$$

converges, then it follows that the expectation of the variable ζ_v exists and is equal to

$$M\zeta_v = \sum_{n=1}^{\infty} p_n A_n$$

where

$$A_n = M\zeta_n = a_1 + a_2 + \dots + a_n$$

Proof. By virtue of the assumptions that have been made

$$\sum_{n=1}^{\infty} p_n A_n = \sum_{n=1}^{\infty} P_n a_n$$

Since ξ_n does not depend on the event $\{v < n\}$, it does not depend on the contrary event $\{v \geq n\}$ either, and for this reason

$$a_n = M\xi_n = M\{\xi_n / v \geq n\}$$

Taking into account the equations that have just been written and also the earlier given properties of conditional expectations,

we can write the following sequence of equalities:

$$\begin{aligned} \sum_{n=1}^{\infty} p_n A_n &= \sum_{n=1}^{\infty} \mathbf{P} \{v \geq n\} \mathbf{M} \{\xi_n/v \geq n\} = \sum_{n=1}^{\infty} \int_{\{v \geq n\}} \xi_n \mathbf{P}(de) = \\ &= \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \int_{\{v=m\}} \xi_n \mathbf{P}(de) \end{aligned}$$

And since the variable $|\xi_n|$ and the event $\{v \geq n\}$ are also independent, it follows that

$$\begin{aligned} \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \left| \int_{\{v=m\}} \xi_n \mathbf{P}(de) \right| &\leq \sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \int_{S_m} |\xi_n| \mathbf{P}(de) = \\ &= \sum_{n=1}^{\infty} \int_{\{v \geq n\}} |\xi_n| \mathbf{P}(de) = \sum_{n=1}^{\infty} \mathbf{P} \{v \geq n\} \mathbf{M} \{|\xi_n|/v \geq n\} = \\ &= \sum_{n=1}^{\infty} \mathbf{P} \{v \geq n\} \mathbf{M} |\xi_n| = \sum_{n=1}^{\infty} P_n c_n < +\infty \end{aligned}$$

The estimate just obtained permits us to write the equation

$$\sum_{n=1}^{\infty} \sum_{m=n}^{\infty} \int_{S_m} \xi_n \mathbf{P}(de) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \int_{S_m} \xi_n \mathbf{P}(de) = \sum_{m=1}^{\infty} \int_{S_m} \zeta_m \mathbf{P}(de)$$

Since

$$\mathbf{M} \zeta_v = \int_U \zeta_v \mathbf{P}(de) = \sum_{m=1}^{\infty} \int_{S_m} \zeta_m \mathbf{P}(de)$$

it follows that the preceding equation proves the theorem.

Corollary. *If in the preceding theorem we put $a = a_1 = a_2 = \dots$, then*

$$\mathbf{M} \zeta_v = a \mathbf{M} v = a \sum_{n=1}^{\infty} n p_n$$

Sec. 30. Moments

The expectation of the variable $(\xi - a)^k$ is called the *moment of the k th order of the random variable ξ* :

$$v_k(a) = \mathbf{M} (\xi - a)^k \quad (1)$$

If $a = 0$, the moment is called the *k th moment about the origin*. It is readily seen that the first moment about the origin is the expectation of the variable ξ .

If $a = \mathbf{M} \xi$, then the moment is called the *central moment*. It is easy to see that the central moment of the first order is zero, while the central moment of the second order is the variance.

We will denote the moments about the origin by the letter v_k , and the central moments by the letter μ_k , the subscript in both cases indicating the order of the moment.

For the second moment $v_2(a)$ we have the obvious equality

$$v_2(a) = M(\xi - a)^2 = M(\xi - M\xi)^2 + (a - M\xi)^2$$

from which we conclude that the second moment $v_k(a)$ has a least value when $a = M\xi$.

There is a simple relationship between the central moments and the moments about the origin. Indeed,

$$\mu_n = M(\xi - M\xi)^n = \sum_{k=0}^n C_n^k (-M\xi)^{n-k} M\xi^k = \sum_{k=0}^n C_n^k (-M\xi)^{n-k} v_k \quad (2)$$

Since $v_1 = M\xi$, it follows that

$$\mu_n = \sum_{k=2}^n (-1)^{n-k} C_n^k v_k v_1^{n-k} + (-1)^{n-1} (n-1) (v_1)^n \quad (3)$$

Let us write down the moment relations for the first four values of n :

$$\begin{aligned} \mu_0 &= 1, \\ \mu_1 &= 0, \\ \mu_2 &= v_2 - v_1^2, \\ \mu_3 &= v_3 - 3v_2 v_1 + 2v_1^3, \\ \mu_4 &= v_4 - 4v_3 v_1 + 6v_2 v_1^2 - 3v_1^4 \end{aligned} \quad (3')$$

These first moments play a particularly important role in statistics.

The quantity

$$m_k = M|\xi - a|^k \quad (4)$$

is called the *absolute moment* of the k th order.

According to Theorem 1 of Sec. 27,

$$v_k(a) = \int (x - a)^k dF(x) \quad (5)$$

Since we agreed that the random variable ξ has expectation only when the integral depicting it converges absolutely, it is clear that the k th moment of the variable ξ exists if and only if the integral

$$\int |x|^k dF_\xi(x)$$

converges. From this remark it follows that if a random variable ξ has a moment of the k th order, it also has moments of all positive orders less than k . Indeed, since for $r < k$, $|x|^k > |x|^r$, if $|x| > 1$,

it follows that

$$\begin{aligned} \int |x|^r dF_\xi(x) &= \int_{|x| \leq 1} |x|^r dF_\xi(x) + \int_{|x| > 1} |x|^r dF_\xi(x) \leq \\ &\leq \int_{|x| \leq 1} |x|^r dF_\xi(x) + \int_{|x| > 1} |x|^k dF_\xi(x) \end{aligned}$$

The first integral on the right-hand side of the inequality is finite by virtue of the finiteness of the limits of integration and the boundedness of the integrand; the second integral converges by assumption.

Example. Find the central and the central absolute moments of a normally distributed random variable:

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(x-a)^2}{2\sigma^2} \right\}$$

We have

$$\mu_k = \frac{1}{\sigma \sqrt{2\pi}} \int (x-a)^k \exp \left\{ -\frac{(x-a)^2}{2\sigma^2} \right\} dx = \frac{\sigma^k}{\sqrt{2\pi}} \int x^k e^{-\frac{x^2}{2}} dx$$

For odd k , since the integrand is odd,

$$\mu_k = 0$$

For even k ,

$$\mu_k = m_k = \sqrt{\frac{2}{\pi}} \sigma^k \int_0^\infty x^k e^{-\frac{x^2}{2}} dx$$

By the substitution $x^2 = 2z$ we reduce the integral to the form

$$\begin{aligned} \mu_k = m_k &= \sqrt{\frac{2}{\pi}} \sigma^k 2^{\frac{k-1}{2}} \int_0^\infty z^{\frac{k-1}{2}} e^{-z} dz = \sqrt{\frac{2}{\pi}} \sigma^k 2^{\frac{k-1}{2}} \Gamma\left(\frac{k+1}{2}\right) = \\ &= \sigma^k (k-1)(k-3) \dots 1 = \sigma^k \frac{k!}{2^{k/2} \left(\frac{k}{2}\right)!} \end{aligned}$$

When k is odd, the absolute moment is

$$\begin{aligned} m_k &= \sqrt{\frac{2}{\pi}} \sigma^k \int_0^\infty x^k e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi}} \sigma^k 2^{\frac{k-1}{2}} \Gamma\left(\frac{k+1}{2}\right) = \\ &= \sqrt{\frac{2}{\pi}} 2^{\frac{k-1}{2}} \left(\frac{k-1}{2}\right)! \sigma^k \end{aligned}$$

The moments of distribution functions cannot be arbitrary quantities. Indeed, no matter what the constants t_0, t_1, \dots, t_n , the

quadratic form

$$J_n = \int \left(\sum_{k=0}^n t_k (x-a)^k \right)^2 dF(x) = \sum_{j=0}^n \sum_{k=0}^n v_{k+j}(a) t_k t_j \geq 0$$

is nonnegative; for this reason, the first $v_j(a)$ should satisfy the following inequalities:

$$\begin{vmatrix} v_0(a) & v_1(a) & \dots & v_k(a) \\ v_1(a) & v_2(a) & \dots & v_{k+1}(a) \\ \dots & \dots & \dots & \dots \\ v_k(a) & v_{k+1}(a) & \dots & v_{2k}(a) \end{vmatrix} \geq 0 \quad (k=0, 1, 2, \dots, n)$$

The absolute moments obey analogous inequalities.

Concerning absolute moments we prove the following theorem.

Theorem. *If a random variable ξ has an absolute moment of order k , then for any t and τ ($0 < t < \tau < k$)*

$$\sqrt[t]{m_t} \leq \sqrt[\tau]{m_\tau} \leq \sqrt[k]{m_k}$$

where

$$m_t = \mathbf{M} |\xi - a|^t$$

and a is any real number.

Proof. First we prove the theorem for the case when t , τ and k are rational numbers. For the sake of definiteness, let

$$t = \frac{p}{q}, \quad \tau = \frac{s}{q}, \quad k = \frac{u}{q}$$

and, by hypothesis,

$$p < s < u$$

Now let r be some positive integer less than u . We consider the nonnegative quadratic form

$$m_{\frac{r-1}{q}} u^2 + 2m_{\frac{r}{q}} uv + m_{\frac{r+1}{q}} v^2 = \int \left[u |x|^{\frac{r-1}{2q}} + v |x|^{\frac{r+1}{2q}} \right]^2 dF(x)$$

The condition that it be nonnegative is, as we know, that

$$m_{\frac{r}{q}}^2 \leq m_{\frac{r-1}{q}} m_{\frac{r+1}{q}}$$

This inequality may obviously also be written as

$$m_{\frac{r}{q}}^{2r} \leq m_{\frac{r-1}{q}}^r m_{\frac{r+1}{q}}^r$$

If we assign to r a succession of values from 1 to r , we get a sequence of inequalities:

$$\begin{aligned} m_{\frac{1}{q}}^2 &\leq m_0 m_{\frac{2}{q}}, \\ m_{\frac{2}{q}}^2 &\leq m_{\frac{1}{q}}^2 m_{\frac{3}{q}}^2, \\ &\dots \dots \dots \\ m_{\frac{r}{q}}^{2r} &\leq m_{\frac{r-1}{q}}^r m_{\frac{r+1}{q}}^r \end{aligned}$$

Note that m_0 is always equal to 1. Then, multiplying these inequalities together and cancelling, we arrive at the inequality

$$m_{\frac{r}{q}}^{r+1} \leq m_{\frac{r+1}{q}}^r$$

Thus,

$$m_{\frac{r}{q}}^{\frac{1}{r}} \leq m_{\frac{r+1}{q}}^{\frac{1}{r+1}}$$

or

$$m_{\frac{r}{q}}^{\frac{q}{r}} \leq m_{\frac{r+1}{q}}^{\frac{a}{r+1}}$$

This inequality quite obviously proves the theorem for the case of t , τ and k being rational.

Since the function m_t is continuous with respect to the argument t in the region $0 \leq t \leq k$, passage to the limit will convince us that the theorem holds for any t , τ and k .

Note that the foregoing theorem contains the following important property of moments:

$$m_1 \leq m_{\frac{1}{2}}^{\frac{1}{2}} \leq m_{\frac{2}{3}}^{\frac{1}{3}} \leq \dots \leq m_k^{\frac{1}{k}} m_{k+1}^{\frac{1}{k+1}} \leq \dots$$

In the examples of the preceding sections, the first two moments of a random variable fully determined its distribution function if the type of function was known beforehand (this occurred in the case of the normal, Poisson, uniform and other distributions). A substantial role in mathematical statistics is played by distribution laws that depend on more than two parameters. If it is known beforehand that a random variable obeys a definite kind of law and only the values of the parameters are unknown, then these parameters are determined in the most important cases by the first moments. But if we do not know the type of distribution function, then, generally speaking, not only a knowledge of the first moments but also of all integral

moments will fail to determine the unknown distribution function. It is possible, it turns out, to construct examples of distribution functions with identical moments of all integral-valued orders. This brings forth the following problem (the *problem of moments*): given a sequence of constants

$$c_0=1, c_1, c_2, c_3, \dots$$

(1) Under what conditions does there exist a distribution function $F(x)$ such that for all n the following equation is valid:

$$c_n = \int x^n dF(x)?$$

(2) When is this function unique?

This problem has been completely solved, but we shall not go into the solution for it would take us outside the scope of this book.

We shall define some more numerical characteristics of random variables that are frequently used in theory and applications.

The *median* of a distribution $F(x)$ is that value of the argument m for which the following inequalities hold:

$$F(m) \leq \frac{1}{2} \leq F(m+0)$$

If $F(x)$ is continuous, there exists at least one m for which the equation

$$F(m)=0.5$$

is valid. If the curve $y=F(x)$ and the straight line $y=0.5$ have a common closed interval, then any point of this interval is the median.

The median exists for all distributions, but the expectation may not.

Note that the median has the following property:

Theorem. *The absolute moment $M|\xi - c|$ for a continuous distribution $F(x)$ assumes a least value if c is chosen equal to the median of the distribution.*

Proof. The theorem follows immediately from the following easily verifiable equality:

$$M|\xi - c| = \begin{cases} M|\xi - m| + 2 \int_m^c (c - x) dF(x) & \text{if } c > m \\ M|\xi - m| + 2 \int_c^m (x - c) dF(x) & \text{if } c < m \end{cases}$$

inasmuch as the second term in both cases is positive for $c \neq m$.

The median of the normal distribution is equal to its mean (expectation).

Just as we defined the median, we define for any number p ($0 < p < 1$) the *distribution quantile of order p* . We confine ourselves to the case of a continuous distribution. Any root of the equation $F(x) = p$ is called the quantile of order p . Clearly, the median is the quantile of order $1/2$. If in a distribution the quantiles are known for a large number of values, say for $p = 0.1; 0.2; \dots; 0.9$ (these quantiles are called *deciles*), they give a sufficiently complete idea about the peculiarities of the distribution.

If a random variable is continuous, i.e. its distribution function has a density, then the value of the argument for which the density is a maximum is called the *mode of the distribution*. For the normal distribution, the mode coincides with the median and the expectation.

Of the other numerical characteristics, most essential are *semi-invariants* (or *cumulants*), which will be defined in Chapter 7. For the present we note the following. In the addition of independent random variables the moment of a sum is, generally speaking, not equal to the sum of the moments of the summands. For the moment of the sum of the independent variables ξ and η we have the equation

$$M(\xi + \eta)^n = \sum_{k=0}^n C_n^k M\xi^k M\eta^{n-k}$$

Cumulants of different orders have the property that upon addition of the independent terms the cumulant of the sum is equal to the sum of the cumulants of summands of the same order. It turns out that the cumulant of any order k is a rational function of the moments of orders less than or equal to k .

EXERCISES

1. A random variable ξ takes on only integral nonnegative values with probabilities

$$(a) \quad P(\xi = k) = \frac{a^k}{(1+a)^{k+1}}, \quad a > 0 \text{ is a constant (this is the } \textit{Pascal distribution}).$$

$$(b) \quad p_k = P\{\xi = k\} = \left(\frac{\alpha\lambda}{1+\alpha\lambda} \right)^k \frac{(1+\alpha) \dots (1+(k-1)\alpha)}{k!} p_0 \text{ for } k > 0 \text{ where } \alpha > 0, \lambda > 0 \text{ and}$$

$$p_0 = P\{\xi = 0\} = (1+\alpha\lambda)^{-\frac{1}{\alpha}}$$

This is the *Polya distribution*.

Find $M\xi$ and $D\xi$.

2. Let μ be the number of occurrences of an event A in n independent trials, in each of which $P(A) = p$. Find

$$(a) \quad M\mu^3, \quad (b) \quad M\mu^4, \quad (c) \quad M|\mu - np|$$

3. The probability that event A will occur in the i th trial is p_i . Let μ be the number of occurrences of A in the first n independent trials. Find

$$(a) M\mu, (b) D\mu, (c) M\left(\mu - \sum_1^n p_i\right)^3 \text{ and } (d) M\left(\mu - \sum_1^n p_i\right)^4$$

4. Prove that, given the conditions of the preceding problem, $D\mu$ reaches a maximum for the given value of $a = \frac{1}{n} \sum_1^n p_i$ provided

$$p_1 = p_2 = \dots = p_n = a$$

5. Let μ be the number of occurrences of an event A in n independent trials, in each of which $P(A) = p$. Also, let a variable η be 0 or 1 depending on whether μ proves to be even or odd. Find $M\eta$.

6. The density function of a random variable ξ is

$$p(x) = \frac{1}{2\alpha} e^{-\frac{|x-a|}{\alpha}}$$

(the Laplace distribution). Find $M\xi$ and $D\xi$.

7. The density function of the absolute speed of a molecule is given by the Maxwell distribution

$$p(x) = \frac{4x^2}{\alpha^3 \sqrt{\pi}} e^{-\frac{x^2}{\alpha^2}} \text{ for } x > 0$$

and $p(x) = 0$ for $x \leq 0$ ($\alpha > 0$ is a constant). Find the average speed of a molecule, its variance, the mean kinetic energy of the molecule (the mass of the molecule is m) and the variance of the kinetic energy.

8. The probability density that a molecule in Brownian motion will be at a distance x from a reflecting wall at time t , if at time t_0 it was at a distance x_0 , is given by the formula

$$p(x) = \begin{cases} \frac{1}{2\sqrt{\pi Dt}} \left\{ e^{-\frac{(x+x_0)^2}{4Dt}} + e^{-\frac{(x-x_0)^2}{4Dt}} \right\} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Find the expectation and the variance of the magnitude displacement of the molecule during the time from $t=t_0$ to t .

9. Prove that for an arbitrary random variable ξ , the possible values of which lie in the interval (a, b) , the following inequalities are valid:

$$a \leq M\xi \leq b \text{ and } D\xi \leq \left(\frac{b-a}{2}\right)^2$$

10. Let x_1, x_2, \dots, x_k be possible values of a random variable ξ . Prove that as $n \rightarrow \infty$

$$(a) \frac{M\xi^{n+1}}{M\xi^n} \rightarrow \max_{1 \leq j \leq k} x_j, (b) \sqrt[n]{M\xi^n} \rightarrow \max_{1 \leq j \leq k} x_j$$

11. Let $F(x)$ be the distribution function of ξ . Prove that if $M\xi$ exists, then

$$M\xi = \int_0^\infty [1 - F(x) + F(-x)] dx$$

and for the existence of $M\xi$ it is necessary that

$$\lim_{x \rightarrow -\infty} xF(x) = \lim_{x \rightarrow \infty} x[1 - F(x)] = 0$$

12. Two points are dropped at random on the line-segment $(0, l)$. Find the expectation, variance and the expectation of the n th power of the distance between them.

13. A random variable ξ is distributed according to the logarithmic normal law; i.e., for $x > 0$ the density function of ξ is

$$p(x) = \frac{1}{x\beta\sqrt{2\pi}} e^{-\frac{1}{2\beta^2}(\ln x - \alpha)^2}$$

($p(x) = 0$ for $x \leq 0$). Find $M\xi$ and $D\xi$.

(A. N. Kolmogorov has demonstrated that particle sizes in crushing obey the logarithmic normal distribution law.)

14. A random variable ξ is normally distributed. Find $M|\xi - a|$ where $a = M\xi$.

15. A box contains 2^n tickets; the number i ($i = 0, 1, \dots, n$) is written on C'_n of them. A total of m tickets are drawn at random, s is the sum of the numbers written on them; find Ms and Ds .

16. The random variables $\xi_1, \xi_2, \dots, \xi_{n+m}$ ($n > m$) are independent, identically distributed and have a finite variance. Find the correlation coefficient of the sums

$$s = \xi_1 + \xi_2 + \dots + \xi_n \quad \text{and} \quad \sigma = \xi_{m+1} + \xi_{m+2} + \dots + \xi_{m+n}$$

17. The random variables ξ and η are independent and are normally distributed with the same parameters a and σ . Find the correlation coefficient of the quantities $\alpha\xi + \beta\eta$ and $\alpha\xi - \beta\eta$, and also their joint distribution.

18. A random vector (ξ, η) is normally distributed; $M\xi = a$, $M\eta = b$, $D\xi = \sigma_1^2$, $D\eta = \sigma_2^2$, and R is the correlation coefficient of ξ and η . Prove that $R = \cos q\pi$ where $q = P\{(\xi - a)(\eta - b) < 0\}$.

19. Let x_1 and x_2 be the results of two independent observations of a normally distributed variable ξ . Prove that $M \max(x_1, x_2) = a + \frac{\sigma}{\sqrt{\pi}}$ where $a = M\xi$, $\sigma^2 = D\xi$.

20. A random vector (ξ, η) is normally distributed, $M\xi = M\eta = 0$, $D\xi = D\eta = 1$, $M\xi\eta = R$. Prove that

$$M \max(\xi, \eta) = \sqrt{\frac{1-R}{\pi}}$$

21. The unevenness in length of cotton fibre is given by

$$\lambda = \frac{a'' - a'}{a}$$

where a is the expectation of fibre length, a'' is the expectation of lengths of the fibres longer than a , and a' is the expectation of lengths of fibres shorter than a . Find the relation between the following quantities:

(a) λ , a , $M|\xi - a|$;

(b) λ , a and σ if ξ is normally distributed.

22. The random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ are independent and uniformly distributed over $(0, 1)$. Let ν be a random variable equal to the k for which the sum

$$s_k = \xi_1 + \xi_2 + \dots + \xi_n$$

exceeds 1 for the first time. Prove that $M\nu = e$.

23. Let ξ be a random variable with density function

$$p_{\xi}(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

Find $M \min(|\xi|, 1)$.

(Problems 22 and 23 were communicated to me by M. I. Yadrenko.)

CHAPTER 6

The Law of Large Numbers

Sec. 31. Mass-Scale Phenomena and the Law of Large Numbers

The vast experience accumulated by mankind teaches us that phenomena which have probability extremely close to unity almost definitely take place. Conversely, events the probability of occurrence of which is very small (close to zero, in other words) occur very infrequently. This circumstance plays a basic role in all practical conclusions from probability theory, for this *experimental fact* enables us in practical activities to consider events that are highly improbable to be *practically impossible*, and events that occur with probabilities close to one as *practically certain* events. And yet we are not able to give an unambiguous answer to the very natural question: what must the probability be so that we can regard an event as practically impossible (practically certain). This is quite natural, since in practical affairs one has to take into account the importance of the events we deal with.

For instance, if in measuring the distance between two villages it were found equal to 5,340 metres and the error of this measurement were equal to or greater than 20 metres with a probability 0.02, we could neglect the possibility of such an error and consider that the distance is indeed equal to 5,340 metres. Thus, in our case we consider the event having probability 0.02 as of practically no importance and disregard it in our practical work. Yet in other cases one cannot neglect probabilities of 0.02 and even less. To illustrate, suppose, in the construction of a large hydroelectric power station that requires enormous expenditure of materials and manpower, it were found that the probability of a catastrophic flood level were equal to 0.02 under the conditions at hand, then this probability would be considered high and it would have to be taken into account in the designing of the station and not neglected, as in the earlier example.

Thus, only the demands of actual practice can suggest the criteria according to which events are to be regarded as practically impossible or practically certain.

At the same time we must note that any event having positive probability, no matter how small, can occur, and if the number of trials in each of which it can occur with one and the same probability is very great, then the probability of at least a single occurrence may be arbitrarily close to unity. This circumstance should be constantly borne in mind. However, if the probability of some event is very small, then it is exceedingly difficult to expect its occurrence in some trial *specified beforehand*. Thus, if somebody asserts that in the first deal of cards between four players each will receive cards of only one suit, then it is natural to suspect that the dealer had certain things in mind, say definite order of the cards known only to him. This confidence is based on the fact that the probability of such a deal, given well shuffled cards, is equal to $(9!)^4/36! < 1.1 \times 10^{-18}$, which is extraordinarily small. Be that as it may, it is on record that cards have been dealt in that way. This instance is a sufficiently good illustration of the difference between the notion of practical impossibility and categorical, so to say, impossibility.

From what has been said it is clear that in our practical activities, and in general theoretical problems as well, events with probabilities close to unity or zero are of great importance. It is thus clear that one of the principal problems of probability theory should be the establishment of regularities involving probabilities close to unity; here, a particular role should be played by laws that arise due to the superimposition of a large number of independent or weakly dependent random factors. The *law of large numbers* is one such proposition of the theory of probability and the most important one.

It would now be natural to regard the law of large numbers as defining the entire assemblage of propositions asserting with probability arbitrarily close to unity that some event will occur that depends on a boundlessly increasing number of random events, each of which exerts on it only a slight effect.

This general conception of theorems akin to the law of large numbers may be formulated somewhat more definitely. Let there be given a sequence of random variables

$$\xi_1, \xi_2, \dots, \xi_n, \dots \quad (1)$$

We consider the random variables ζ_n which are certain specified symmetric functions of the first n variables of the sequence (1):

$$\zeta_n = f_n(\xi_1, \xi_2, \dots, \xi_n)$$

If there exists a sequence of constants $a_1, a_2, \dots, a_n, \dots$ such that for any $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbf{P} \{ |\zeta_n - a_n| < \varepsilon \} = 1 \quad (2)$$

then the sequence (1) obeys the law of large numbers with given functions f_n .

Ordinarily, however, a much more definite meaning is ascribed to the concept of the law of large numbers. Namely, we restrict ourselves to the case when f_n is the *arithmetic mean* of the variables $\xi_1, \xi_2, \dots, \xi_n$.

If all the quantities a_n in relation (2) are equal to one and the same quantity a , then we say that the random variables ξ_n *converge in probability* to a . In these terms, relation (2) means that $\xi_n - a_n$ converges in probability to zero.

When studying single phenomena, we observe them together with all their individual peculiarities that obscure the manifestation of laws involved in the observation of large numbers of similar phenomena. It was noticed a long time ago that factors which are not connected with the essence of the process as a whole and which appear only in single instances mutually cancel out when one considers the average of a large number of observations.

Later, this empirical result was noted with increasing frequency, yet as a rule no attempt was ever made to give any theoretical explanation. Incidentally, for many authors no explanation was required, since the presence of regularities both in natural and social phenomena was, to them, nothing other than a manifestation of the rules of divine order.

Even today some authors impoverish the content of the law of large numbers and even distort its methodological significance by simply reducing it to an experimentally observed regularity. Actually, the enduring scientific value of the investigations of Chebyshev, Markov and other researchers in the field of the law of large numbers does not consist in the fact that they detected the empirical stability of means, but in the fact that they found the general conditions whose fulfillment definitely brings about the statistical stability of means.

To illustrate the operation of the law of large numbers, we take the following schematic example. According to modern physical views, a gas consists of an enormous number of individual particles in constant and chaotic motion. Speaking of each separate molecule, one cannot predict the velocity it will have and the place it will be in at any given instant of time. However, we can, given certain conditions of the gas, calculate the portion of molecules that will be moving with a given velocity or the portion of them that will be located in a given volume. But, strictly speaking, that is precisely what the physicist wishes to know, since the basic characteristics of a gas—pressure, temperature, viscosity, and so forth—are determined not by the bizarre behaviour of a single molecule but by the collective action of all of them. Thus, the pressure of a gas is equal to the overall action of molecules impinging on a plate of unit area in unit time. The number of impacts and the speeds of the impinging

molecules vary according to chance; however, by virtue of the law of large numbers (in Chebyshev's form) the pressure should be nearly constant. This "equalizing" effect of the law of large numbers in physical phenomena is exhibited with exceptional exactitude. Suffice it to recall that, say, under ordinary conditions even very precise measurements hardly at all permit noticing deviations from Pascal's law of the pressure of a liquid. This extraordinary agreement of theory and experiment even served the opponents of the molecular structure of matter with a peculiar kind of argument: if matter were molecular in structure, then departures from Pascal's law would be in evidence. Such deviations, the so-called fluctuations of pressure, were actually observed when scientists had learned to isolate relatively small quantities of molecules, as a result of which the effect of separate molecules did not completely even out and remained rather strong.

Sec. 32. Chebyshev's Form of the Law of Large Numbers

We shall now formulate and prove the theorems of Chebyshev, Markov and others. The method used in this case belongs to Chebyshev.

Chebyshev's Inequality. *For every random variable ξ having a finite variance, the inequality*

$$P \{ |\xi - M\xi| \geq \varepsilon \} \leq \frac{D\xi}{\varepsilon^2} \quad (1)$$

is valid for every $\varepsilon > 0$.

Proof. If $F(x)$ denotes the distribution function of the random variable ξ , then

$$P \{ |\xi - M\xi| \geq \varepsilon \} = \int_{|x - M\xi| \geq \varepsilon} dF(x)$$

Since in the region of integration

$$\frac{|x - M\xi|}{\varepsilon} \geq 1$$

it follows that

$$\int_{|x - M\xi| \geq \varepsilon} dF(x) \leq \frac{1}{\varepsilon^2} \int_{|x - M\xi| \geq \varepsilon} (x - M\xi)^2 dF(x)$$

But we only strengthen this inequality by extending the integration to all values of x :

$$\int_{|x - M\xi| \geq \varepsilon} dF(x) \leq \frac{1}{\varepsilon^2} \int (x - M\xi)^2 dF(x) = \frac{D\xi}{\varepsilon^2}$$

This completes the proof of Chebyshev's inequality.

Chebyshev's Theorem. If $\xi_1, \xi_2, \dots, \xi_n, \dots$ are a sequence of pairwise independent random variables having finite variances, bounded by one and the same constant

$$D\xi_1 \leq C, \quad D\xi_2 \leq C, \quad \dots, \quad D\xi_n \leq C, \quad \dots$$

then, for any constant $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n M\xi_k \right| < \varepsilon \right\} = 1 \quad (2)$$

Proof. We know that by hypothesis

$$D \left(\frac{1}{n} \sum_{k=1}^n \xi_k \right) = \frac{1}{n^2} \sum_{k=1}^n D\xi_k$$

and, consequently,

$$D \left(\frac{1}{n} \sum_{k=1}^n \xi_k \right) \leq \frac{C}{n}$$

According to the Chebyshev inequality,

$$\mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n M\xi_k \right| < \varepsilon \right\} \geq 1 - \frac{D \left(\frac{1}{n} \sum_{k=1}^n \xi_k \right)}{\varepsilon^2} \geq 1 - \frac{C}{n\varepsilon^2}$$

Passing to the limit as $n \rightarrow \infty$, we get

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n M\xi_k \right| < \varepsilon \right\} \geq 1$$

And since probability cannot exceed one, the theorem thus follows.

We shall take note of certain important special cases of Chebyshev's theorem.

1. Bernoulli's Theorem. Let μ be the number of occurrences of an event A in n independent trials and p the probability of occurrence of event A in each of the trials. Then, for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| < \varepsilon \right\} = 1 \quad (3)$$

Proof. Indeed, by introducing the random variables μ_k , equal to the number of occurrences of A in the k th trial, we have

$$\mu = \mu_1 + \mu_2 + \dots + \mu_n$$

And since

$$M\mu_k = p, \quad D\mu_k = pq \leq \frac{1}{4}$$

it follows that the Bernoulli theorem is an elementary special case of the Chebyshev theorem.

Since in practical work it is frequently necessary to determine unknown probabilities in approximate fashion experimentally, agreement between the Bernoulli theorem and experiment has been verified by performing large numbers of experiments. Here, events were considered in which the probabilities may, for one reason or another, be regarded as known, and concerning which it was easy to perform trials and ensure the independence of the trials as well as the constancy of the probabilities in each of the trials. All such experiments yielded excellent agreement with theory. We indicate the results of some of these readily reproducible experiments.

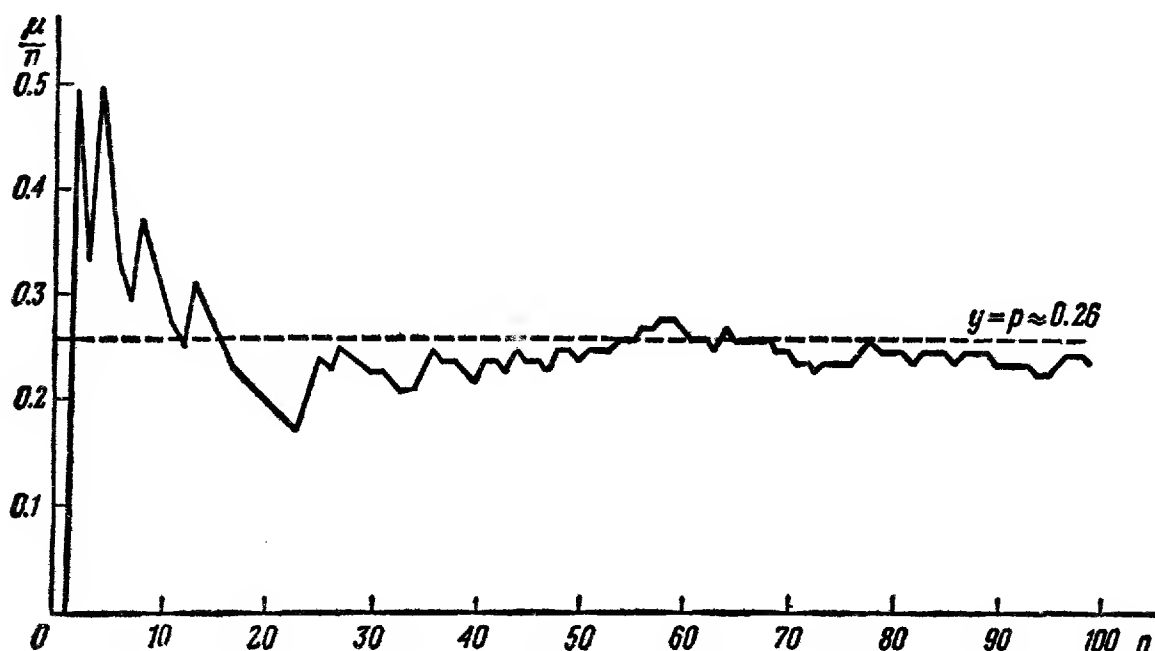


Fig. 20

A deck of 36 cards was divided in half at random 100 times. The results are tabulated in Table 11. The first column indicates the number of the trial, the second, the number of red cards in half the deck, the third, the number of cases in which red and black cards came out half and half in the trials, and, finally, the fourth column gives the frequencies.

In Example 3, Sec. 5, it was computed that the probability of obtaining equal numbers of black and red cards in each half deck is

$$p = \frac{(18!)^4}{30! (9!)^4} \approx 0.26$$

The curve in Fig. 20 gives a clear-cut idea of the variation of the frequency $\frac{\mu}{n}$ as a function of the number of trials. At first, when the number of experiments is small, the broken line sometimes departs

appreciably from the straight line $y=p \approx 0.26$. Then, as the number of experiments increases, the broken line, on the whole, comes closer and closer to the straight line.

In the case at hand, the result was a rather considerable final (for $n=100$) deviation of frequency from the probability (roughly equal to 0.02). By the Laplace theorem, the probability of obtaining such a deviation or a greater one is equal to

$$\begin{aligned} \mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| \geq 0.02 \right\} &= \\ &= \mathbf{P} \left\{ \left| \frac{\mu - np}{\sqrt{npq}} \right| \geq 0.02 \sqrt{\frac{n}{pq}} \right\} \approx 1 - 2\Phi \left(0.02 \sqrt{\frac{n}{pq}} \right) = \\ &= 1 - 2\Phi \left(0.02 \sqrt{\frac{100}{0.26 \cdot 0.74}} \right) = 1 - 2\Phi(0.455) \sim 0.65 \end{aligned}$$

Thus, if the experiment is repeated a large number of times, roughly in two thirds of the cases the deviation will not be less than what we obtained in our experiment.

The eighteenth century French naturalist Buffon tossed a coin 4040 times with heads appearing 2048 times. In Buffon's experiment, the frequency of heads turning up is approximately equal to 0.507.

The English statistician Karl Pearson tossed a coin 12,000 times and obtained heads 6,019 times. The frequency of heads in Pearson's experiment is 0.5016.

On another occasion he threw a coin 24,000 times obtaining heads 12,012 times, the frequency of occurrence of heads being 0.5005. In all these experiments, the frequency only slightly deviated from the probability—0.5.

2. Poisson's Theorem. *If in a sequence of independent trials, the probability of occurrence of an event A in the k th trial is equal to p_k , then*

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{\mu}{n} - \frac{p_1 + p_2 + \dots + p_n}{n} \right| < \varepsilon \right\} = 1$$

where, as usual, μ denotes the number of occurrences of A in the first n trials.

Introducing the random variables μ_k , which are equal to the number of occurrences of A in the k th trial, and noting that

$$\mathbf{M}\mu_k = p_k, \quad \mathbf{D}\mu_k = p_k q_k \leq \frac{1}{4}$$

we find that the Poisson theorem is a special case of the Chebyshev theorem.

TABLE 11

Trial No.	Number of red cards	Number of favourable cases	Frequency	Trial No.	Number of red cards	Number of favourable cases	Frequency
1	8	0	0.00	51	9	13	0.25
2	9	1	0.50	52	8	13	0.25
3	11	1	0.33	53	7	13	0.25
4	9	2	0.50	54	9	14	0.26
5	11	2	0.40	55	7	14	0.26
6	8	2	0.33	56	9	15	0.27
7	11	2	0.29	57	9	16	0.28
8	9	3	0.37	58	11	16	0.28
9	8	3	0.33	59	8	16	0.27
10	7	3	0.30	60	8	16	0.27
11	12	3	0.27	61	8	16	0.26
12	10	3	0.25	62	10	16	0.26
13	9	4	0.31	63	12	16	0.25
14	13	4	0.29	64	9	17	0.27
15	12	4	0.27	65	11	17	0.26
16	8	4	0.25	66	12	17	0.26
17	11	4	0.23	67	11	17	0.26
18	10	4	0.22	68	8	17	0.25
19	8	4	0.21	69	10	17	0.25
20	11	4	0.20	70	8	17	0.25
21	12	4	0.19	71	7	17	0.24
22	10	4	0.18	72	9	18	0.25
23	10	4	0.17	73	10	18	0.25
24	9	5	0.21	74	8	18	0.24
25	9	6	0.24	75	11	18	0.24
26	14	6	0.23	76	8	18	0.24
27	9	7	0.26	77	9	19	0.25
28	10	7	0.25	78	9	20	0.26
29	10	7	0.24	79	5	20	0.26
30	7	7	0.23	80	8	20	0.25
31	10	7	0.22	81	7	20	0.25
32	7	7	0.22	82	10	20	0.24
33	8	7	0.21	83	9	21	0.25
34	10	7	0.21	84	6	21	0.25
35	9	8	0.23	85	10	21	0.25
36	9	9	0.25	86	10	21	0.24
37	10	9	0.24	87	9	22	0.25
38	10	9	0.24	88	7	22	0.25
39	8	9	0.23	89	7	22	0.25
40	7	9	0.22	90	10	22	0.24
41	9	10	0.24	91	8	22	0.24
42	10	10	0.24	92	8	22	0.24
43	10	10	0.23	93	10	22	0.24
44	9	11	0.25	94	8	22	0.23
45	8	11	0.24	95	11	22	0.23
46	7	11	0.24	96	9	23	0.24
47	12	11	0.23	97	9	24	0.25
48	9	12	0.25	98	10	24	0.25
49	6	12	0.25	99	7	24	0.24
50	7	12	0.24	100	7	24	0.24

3. If a sequence of pairwise independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ is such that

$$M\xi_1 = M\xi_2 = \dots = M\xi_n = \dots = a$$

and

$$D\xi_1 \leq C, D\xi_2 \leq C, \dots, D\xi_n \leq C, \dots$$

then for any constant $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - a \right| < \varepsilon \right\} = 1$$

This special case of the Chebyshev theorem serves as a basis for the rule of the arithmetic mean that is constantly employed in the theory of measurements. Suppose we are measuring a certain physical quantity a . Repeating the measurement n times under identical conditions, the observer will obtain results x_1, x_2, \dots, x_n that do not exactly coincide. The rule is to obtain an approximate value of a by taking the arithmetic mean of the observational results:

$$a \approx \frac{x_1 + x_2 + \dots + x_n}{n}$$

If the measurements do not exhibit a systematic error, that is, if

$$Mx_1 = Mx_2 = \dots = Mx_n = a$$

and if there is no uncertainty about the observed values themselves, then according to the law of large numbers, for sufficiently large values of n with a probability arbitrarily close to one we can in this way obtain a value that is arbitrarily close to the desired quantity a .

The Chebyshev inequality permits obtaining a stronger result in the case of identically distributed independent terms.

Khinchin's Theorem. If the random variables ξ_1, ξ_2, \dots are independent and identically distributed with finite expectations ($a = M\xi_n$), then as $n \rightarrow \infty$

$$P \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - a \right| < \varepsilon \right\} \rightarrow 1$$

Proof. We take advantage of a device which was first employed by A. A. Markov in 1907 and later became known as the *method of truncation*. This procedure is frequently used in modern probability theory.

Define the new random variables by the following rule: let $\delta > 0$ be fixed and for $k = 1, 2, \dots, n$

$$\begin{aligned} \eta_k &= \xi_k, & \zeta_k &= 0, & \text{if } |\xi_k| < \delta n \\ \eta_k &= 0, & \zeta_k &= \xi_k, & \text{if } |\xi_k| \geq \delta n \end{aligned}$$

It is obvious that for any k ($1 \leq k \leq n$)

$$\xi_k = \eta_k + \zeta_k$$

For the variables η_k there exist expectation and variance

$$\begin{aligned} a_n &= M\eta_k = \int_{-\delta n}^{\delta n} x dF(x) \\ D\eta_k &= \int_{-\delta n}^{\delta n} x^2 dF(x) - a_n^2 \leq \delta n \int_{-\delta n}^{\delta n} |x| dF(x) \leq \delta b n \end{aligned}$$

where $b = \int_{-\infty}^{\infty} |x| dF(x)$. Since as $n \rightarrow \infty$

$$a_n \rightarrow a$$

it follows that for any $\varepsilon > 0$, given sufficiently large n ,

$$|a_n - a| < \varepsilon \quad (4)$$

By virtue of Chebyshev's inequality

$$\mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \eta_k - a_n \right| \geq \varepsilon \right\} \leq \frac{b\delta}{\varepsilon^2}$$

Now using inequality (4) we get:

$$\mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \eta_k - a \right| \geq 2\varepsilon \right\} \leq \frac{b\delta}{\varepsilon^2}$$

Now note that

$$\mathbf{P} \{ \zeta_n \neq 0 \} = \int_{|x| \geq \delta n} dF(x) \leq \frac{1}{\delta n} \int_{|x| \geq \delta n} |x| dF(x)$$

Since expectation exists, the right-hand side becomes less than $\frac{\delta}{n}$ for sufficiently large n . But

$$\mathbf{P} \left\{ \sum_{k=1}^n \zeta_k \neq 0 \right\} \leq \sum_{k=1}^n \mathbf{P} \{ \zeta_k \neq 0 \} \leq \delta$$

and so

$$\begin{aligned} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - a \right| \geq 2\varepsilon \right\} &\leq \\ &\leq \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \eta_k - a \right| \geq 2\varepsilon \right\} + \mathbf{P} \left\{ \sum_{k=1}^n \zeta_k \neq 0 \right\} \leq \frac{b\delta}{\varepsilon^2} + \delta \end{aligned}$$

Since ε and δ are arbitrary, the right-hand side may be made less than any number; this proves the theorem.

We also formulate Markov's theorem; its proof is an obvious consequence of Chebyshev's inequality.

Markov's Theorem. *If a sequence of random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ is such that as $n \rightarrow \infty$*

$$\frac{1}{n^2} \mathbf{D} \left(\sum_{k=1}^n \xi_k \right) \rightarrow 0 \quad (5)$$

then, for any positive constant ε ,

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n \mathbf{M}\xi_k \right| < \varepsilon \right\} = 1$$

If the random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ are pairwise independent, the Markov condition becomes (as $n \rightarrow \infty$):

$$\frac{1}{n^2} \sum_{k=1}^n \mathbf{D}\xi_k \rightarrow 0$$

From this it is evident that the Chebyshev theorem is a special case of Markov's theorem.

We obtain the following theorem as a direct consequence of Markov's theorem. It was also proved by Markov.

Theorem. *Let μ be the number of occurrences of an event E in n trials connected in a homogeneous Markov chain, and let p_1, p_2, \dots be the probabilities of occurrence of E in the first, second, and so forth trials, respectively; then, for any $\varepsilon > 0$,*

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{\mu}{n} - \frac{1}{n} \sum_{k=1}^n p_k \right| < \varepsilon \right\} = 1 \quad (6)$$

The proof of the theorem is obvious by virtue of the results of Example 6 in Sec. 28.

Since according to the results of this example

$$\frac{1}{n} \sum_{k=1}^n p_k = p + o(1)$$

it follows that (6) is equivalent to the equation

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{\mu}{n} - p \right| < \varepsilon \right\} = 1$$

In this form, the foregoing theorem is completely analogous to Bernoulli's theorem.

Sec. 33. A Necessary and Sufficient Condition for the Law of Large Numbers

We have already pointed out that the law of large numbers is one of the basic propositions of probability theory. This makes it clear why so much effort has gone into establishing the broadest possible conditions to be satisfied by the variables $\xi_1, \xi_2, \dots, \xi_n, \dots$, so that the law of large numbers should hold.

The history of the problem is as follows. At the end of the 17th century and the beginning of the 18th, James Bernoulli proved a theorem that bears his name. This theorem of Bernoulli was first published in 1713, after the author's death, in the treatise *Ars conjectandi* (the art of constructing conjectures). Then at the beginning of the 19th century, Poisson proved a similar theorem under more general conditions. No further advances were made up to the middle of the 19th century. In 1866 the great Russian mathematician P. L. Chebyshev discovered a method that we have given in Sec. 32. Later, A. A. Markov noticed that Chebyshev's reasoning permits of a still more general result (see Sec. 32).

Further efforts to attain fundamental advances failed until 1926, when A. N. Kolmogorov obtained conditions necessary and sufficient for a sequence of mutually independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ to obey the law of large numbers. In 1928, A. Ya. Khinchin demonstrated that if the random variables ξ_n are not only independent but are also identically distributed, then the existence of the expectation $\mathbf{M}\xi_n$ is a sufficient condition for applying the law of large numbers.

During recent years many papers have been devoted to determining the conditions to be imposed on dependent variables so that the law of large numbers may be applicable. The Markov theorem is a proposition of this type.

By employing the Chebyshev method it is easy to obtain a condition similar to Markov's condition, but this time not only sufficient but also necessary for the applicability of the law of large numbers to a sequence of arbitrary random variables.

Theorem. *So that a sequence of random variables*

$$\xi_1, \xi_2, \xi_3, \dots$$

(arbitrarily dependent) should, for any positive ε , satisfy the relation

$$\lim_{n \rightarrow \infty} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n \mathbf{M} \xi_k \right| < \varepsilon \right\} = 1 \quad (1)$$

it is necessary and sufficient that as $n \rightarrow \infty$

$$\mathbf{M} \frac{\left(\sum_{k=1}^n (\xi_k - \mathbf{M} \xi_k) \right)^2}{n^2 + \left(\sum_{k=1}^n (\xi_k - \mathbf{M} \xi_k) \right)^2} \rightarrow 0 \quad (2)$$

Proof. First suppose that (2) is satisfied; we will show that in this case (1) will also be satisfied. Denote by $\Phi_n(x)$ the distribution function of the variable

$$\eta_n = \frac{1}{n} \sum_{k=1}^n (\xi_k - \mathbf{M} \xi_k)$$

It is easy to verify the following chain of relations

$$\begin{aligned} \mathbf{P} \left\{ \left| \frac{1}{n} \sum_{k=1}^n (\xi_k - \mathbf{M} \xi_k) \right| \geq \varepsilon \right\} &= \mathbf{P} \{ |\eta_n| \geq \varepsilon \} = \\ &= \int_{|x| \geq \varepsilon} d\Phi_n(x) \leq \frac{1+\varepsilon^2}{\varepsilon^2} \int_{|x| \geq \varepsilon} \frac{x^2}{1+x^2} d\Phi_n(x) \leq \\ &\leq \frac{1+\varepsilon^2}{\varepsilon^2} \int \frac{x^2}{1+x^2} d\Phi_n(x) = \frac{1+\varepsilon^2}{\varepsilon^2} \mathbf{M} \frac{\eta_n^2}{1+\eta_n^2}^* \end{aligned}$$

This inequality proves that the conditions of the theorem are sufficient.

* We can write this equality on the basis of the formula

$$\mathbf{M} f(\xi) = \int f(x) dF_\xi(x)$$

(see Theorem 1, Sec. 27).

Let us now show that condition (2) is necessary. It will readily be seen that

$$\begin{aligned}
 \mathbf{P} \{ |\eta_n| \geq \varepsilon \} &= \int_{|x| \geq \varepsilon} d\Phi_n(x) \geq \int_{|x| \geq \varepsilon} \frac{x^2}{1+x^2} d\Phi_n(x) = \\
 &= \int \frac{x^2}{1+x^2} d\Phi_n(x) - \int_{|x| < \varepsilon} \frac{x^2}{1+x^2} d\Phi_n(x) \geq \\
 &\geq \int \frac{x^2}{1+x^2} d\Phi_n(x) - \varepsilon^2 = \mathbf{M} \frac{\eta_n^2}{1+\eta_n^2} - \varepsilon^2 \quad (3)
 \end{aligned}$$

Thus,

$$0 \leq \mathbf{M} \frac{\eta_n^2}{1+\eta_n^2} \leq \varepsilon^2 + \mathbf{P} \{ |\eta_n| \geq \varepsilon \}$$

First choose ε sufficiently small and then n sufficiently large; by doing so we can make the right-hand side of the last inequality as small as desired.

We note that all the theorems that were proved in the preceding section follow readily from the general proposition that has just been proved. Indeed, since for any n and any ξ_k the inequality

$$\frac{\eta_n^2}{1+\eta_n^2} \leq \eta^2 = \left[\frac{1}{n} \sum_{k=1}^n (\xi_k - \mathbf{M}\xi_k) \right]^2$$

is valid, it follows that if variances exist we get the following inequality:

$$\mathbf{M} \frac{\eta_n^2}{1+\eta_n^2} \leq \frac{1}{n^2} \mathbf{D} \sum_{k=1}^n \xi_k$$

Thus, if the Markov condition is satisfied, then condition (2) is also satisfied and, consequently, the sequence $\xi_1, \xi_2, \dots, \xi_n, \dots$ obeys the law of large numbers.

Still, we must take note of the fact that in more complicated cases when the variables ξ_k are not assumed to have finite variances, the theorem just proved is of extremely slight use in any real verification of the applicability of the law of large numbers, for condition (2) refers not to the separate summands but to their sums. However, it is apparently impossible to hope to find necessary and sufficient conditions (and what is more, such as to be convenient for applications) without making any assumption about the variables ξ_k and the relationship existing between them.

Attempts at a practical employment of the theorems that have just been proved come up against one fundamental difficulty: can we

take it that the phenomenon or the production process under study proceeds via the action of independent causes? Is there not a contradiction between the very concept of independence and our basic ideas of the interrelationship of phenomena in the external world? In a mathematical study of the phenomena of nature, technical processes or social phenomena we must first of all derive our basic premises on the basis of a profound study of the essence of the phenomenon at hand and its qualitative peculiarities. We have to take into account changes in the external conditions in which our phenomenon develops and alter the mathematical apparatus and the premises underlying its applications as soon as it is seen that the conditions of realization of the phenomenon have changed.

As a first approximation to reality we can assume that the causes operating on the phenomenon are independent, and we can draw conclusions from this supposition. We can judge about how successful our scheme of the phenomenon is and how suitable the mathematical apparatus for its study is by agreement between the theory we have constructed and practice. If our theoretical results depart substantially from experiment, then we will have to revise the premises; in particular, if it is a question of applying the law of large numbers, we may have to give up the supposition of total independence of the operating causes and presume them to be dependent, albeit of a weak nature.

We have already spoken of the fact that accumulated experience in the use of the theorems of the law of large numbers indicates that the condition of independence is satisfactory in many important problems of natural science and technology.

Sec. 34. The Strong Law of Large Numbers

It often happens that the definitely unjustified conclusion is drawn from Bernoulli's theorem that the frequency of an event A tends to the probability of A in the case of a limitless increase in the number of trials. Actually, Bernoulli's theorem states that for a sufficiently large number of trials n the probability of one single inequality

$$\left| \frac{\mu}{n} - p \right| < \varepsilon$$

becomes greater than $1 - \eta$ for an arbitrary $\eta > 0$. In 1909 the French mathematician E. Borel detected a more profound proposition which became known as the *strong law of large numbers*. The formulation and proof of the Borel theorem and also of the more general propositions of Kolmogorov requires the introduction of an important concept: *the convergence of a sequence of random variables*.

Let there be a sequence of random variables defined on one and the same set of elementary events U :

$$\xi_n = f_n(e) \quad (e \in U) \quad (1)$$

We consider the set A of all the elementary events e for which the sequence $f_n(e)$ converges. Let $f(e)$ denote the limit of $f_n(e)$ at the point e . If we let A_{nk}^r denote the set of those e for which the inequality

$$|f_{n+k}(e) - f(e)| < \frac{1}{r} \quad (2)$$

is fulfilled, then it is obvious that

$$A = \prod_{r=1}^{\infty} \sum_{n=1}^{\infty} \prod_{k=1}^{\infty} A_{nk}^r \quad (3)$$

Indeed, if the sequence of functions $f_n(e)$ converges at the point e , then: (1) inequalities (2) must be satisfied for all k if n is sufficiently great; (2) they must be satisfied beginning with a certain n ; (3) they must be satisfied for sufficiently large n for any value of r . Equation (3) states these three requirements symbolically. In accordance with Equation (3) and the definition of a random event, the subset A belongs to the field of random events. We define a random variable ξ as follows: if $e \in A$, then $\xi = f(e)$, but if $e \in \bar{A}$, then $\xi = 0$.

If the probability of a random event A is equal to 1, then we say that the sequence of random variables ξ_n converges to the random variable ξ almost certainly (or that it converges with probability one*).

If the sequence ξ_n converges to ξ almost certainly, we then write this fact as follows:

$$P\{\xi_n \rightarrow \xi\} = 1 \quad (4)$$

This equation can clearly be written differently:

$$P\{\xi_n \not\rightarrow \xi\} = 0 \quad (4')$$

The latter expression signifies that the probability that there will be a number r such that for all n and at least for one value of

* The concept of almost certain convergence corresponds exactly to that of convergence almost everywhere in the theory of functions.

In probability theory a big role is also played by the so-called convergence in probability: a sequence of random variables ξ_n converges in probability to the random variable ξ if, for any $\varepsilon > 0$, the probability of the inequality $|\xi_n - \xi| < \varepsilon$ tends to unity as $n \rightarrow \infty$.

Convergence in probability is an analogue of the convergence in measure of a sequence of functions in the theory of functions.

It is obvious that the law of large numbers asserts that under certain circumstances the sums $\frac{1}{n} \sum_{k=1}^n (\xi_k - M_{\xi_k})$ converge in probability to zero.

$k = k(n)$ the inequality

$$|\xi_{n+k} - \xi| \geq \frac{1}{r}$$

holds is equal to zero.

We now indicate a sufficient condition for the convergence of a sequence of random variables with probability one.

Lemma. *If for any positive integer r ,*

$$\sum_{n=1}^{\infty} \mathbf{P} \left\{ |\xi_n - \xi| \geq \frac{1}{r} \right\} < +\infty \quad (5)$$

then (4) or —what is the same thing—(4') holds.

Proof. Let E'_n denote an event which consists in the fact that the inequality

$$|\xi_n - \xi| \geq \frac{1}{r}$$

is valid.

We further assume

$$S'_n = \sum_{k=1}^{\infty} E'_{n+k}$$

From the fact that

$$\mathbf{P}\{S'_n\} \leq \sum_{k=1}^{\infty} \mathbf{P}\{E'_{n+k}\} = \sum_{l=n+1}^{\infty} \mathbf{P}\left\{|\xi_l - \xi| \geq \frac{1}{r}\right\}$$

we derive, by virtue of (5), the equation

$$\lim_{n \rightarrow \infty} \mathbf{P}\{S'_n\} = 0 \quad (6)$$

Now let

$$S' = S'_1 S'_2 S'_3 \dots$$

From the fact that the event S' implies any one of the events S'_n , we get, by virtue of (6),

$$\mathbf{P}(S') = 0 \quad (7)$$

Finally, set

$$S = S^1 + S^2 + S^3 + \dots$$

It is easy to establish that this event signifies that any r will be found such that for every $n (n = 1, 2, 3, \dots)$ the inequalities

$$|\xi_{n+k} - \xi| \geq \frac{1}{r}$$

will be satisfied for at least one k [$k = k(n)$]. Since

$$P(S) \leq \sum_{r=1}^{\infty} P\{S^r\}$$

it follows that by (7)

$$P\{S\} = 0$$

and this completes the proof.

Repeating word for word the foregoing reasoning, we can obtain a somewhat stronger proposition:

If there exists a sequence of integers $1 = n_1 < n_2 < n_3 < \dots$ such that the series

$$\sum_{k=1}^{\infty} P\left\{\max_{n_k \leq n < n_{k+1}} |\xi_n - \xi| \geq \frac{1}{r}\right\}$$

converges for every positive integer r , then the sequence of random variables ξ_1, ξ_2, \dots almost certainly converges to ξ .

Let us now apply the newly introduced concept and the lemma that we have proved.

Borel's Theorem. *Let μ be the number of occurrences of an event A in n independent trials, in each of which A may occur with probability p . Then, as $n \rightarrow \infty$,*

$$P\left\{\frac{\mu}{n} \rightarrow p\right\} = 1$$

Proof. According to Lemma 1, it suffices to detect convergence of the series

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{\mu}{n} - p\right| \geq \frac{1}{r}\right\} \quad (8)$$

for any natural number r . For this purpose we note that in the very same way that Chebyshev's inequality was proved (Sec. 32) it is possible to establish the following inequality: for every random variable for which $M(\xi - M\xi)^4$ exists,

$$P\{|\xi - M\xi| \geq \varepsilon\} \leq \frac{1}{\varepsilon^4} M(\xi - M\xi)^4$$

Thus,

$$P\left\{\left|\frac{\mu}{n} - p\right| \geq \frac{1}{r}\right\} \leq r^4 M\left(\frac{\mu}{n} - p\right)^4$$

As we have repeatedly done, let us introduce the auxiliary random variables μ_n , equal to the number of occurrences of event A in the

i th trial. Since

$$\frac{\mu}{n} - p = \frac{1}{n} \sum_{i=1}^n (\mu_i - p)$$

it follows that

$$\mathbf{M} \left(\frac{\mu}{n} - p \right)^4 = \frac{1}{n^4} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \mathbf{M}(\mu_i - p)(\mu_j - p)(\mu_k - p)(\mu_l - p) \quad (9)$$

Since $\mathbf{M}(\mu_i - p) = 0$, all the terms having at least one of the factors $(\mu_i - p)$ to the first power vanish. Therefore, in this sum only terms of the form $\mathbf{M}(\mu_i - p)^4$ and $\mathbf{M}(\mu_i - p)^2(\mu_s - p)^2$ are different from zero. It is clear that

$$\mathbf{M}(\mu_i - p)^4 = pq(p^3 + q^3)$$

and

$$\mathbf{M}(\mu_i - p)^2(\mu_s - p)^2 = p^2q^2 \quad (i \neq s)$$

The number of summands of the first type is n , the number of the second type is $3n(n-1)$. Indeed, i may coincide either with j or with k or with l and then take on one of the n values from 1 to n ; s can assume only one of the $n-1$ values because $s \neq i$.

Thus,

$$\mathbf{M} \left(\frac{\mu}{n} - p \right)^4 = \frac{pq}{n^4} [n(p^3 + q^3) + 3pq(n^2 - n)] < \frac{1}{4n^2}$$

and, consequently, the series (8) converges. The proof of the theorem is complete.

Borel's theorem sparked off a whole cycle of investigations devoted to seeking the conditions under which the so-called strong law of large numbers holds.

We say that a *sequence of random variables*

$$\xi_1, \xi_2, \xi_3, \dots$$

obeys the strong law of large numbers if with probability one, as $n \rightarrow \infty$,

$$\frac{1}{n} \sum_{k=1}^n \xi_k - \frac{1}{n} \sum_{k=1}^n \mathbf{M}\xi_k \rightarrow 0$$

Broad yet simple sufficient conditions for accomplishment of the strong law of large numbers are given by a theorem of A. N. Kolmogorov whose proof is based on an interesting generalization of Chebyshev's inequality.

Kolmogorov's Inequality. *If the mutually independent random variables $\xi_1, \xi_2, \dots, \xi_n$ have finite variances, then the probability of*

the simultaneous realization of the inequalities

$$\left| \sum_{s=1}^k (\xi_s - \mathbf{M}\xi_s) \right| < \varepsilon \quad (k=1, 2, \dots, n)$$

is not less than

$$1 - \frac{1}{\varepsilon^2} \sum_{k=1}^n \mathbf{D}\xi_k$$

Proof. We introduce the notations

$$\eta_k = \xi_k - \mathbf{M}\xi_k, \quad S_k = \sum_{j=1}^k \eta_j$$

Also, let E_k denote the event that

$$|S_j| < \varepsilon \text{ for } j \leq k-1 \text{ and } |S_k| \geq \varepsilon \quad (10)$$

E_0 denotes the event that $|S_j| < \varepsilon$ for $j \leq n$.

Since the event that consists in the fact that at least for one k ($1 \leq k \leq n$) the inequality

$$|S_k| \geq \varepsilon \quad (k=1, 2, \dots, n)$$

will hold (in other words, that $\max |S_k| \geq \varepsilon$) is equivalent to the event $\sum_{k=1}^n E_k$, it follows by virtue of the incompatibility of the events E_k that

$$\mathbf{P} \left\{ \max_{1 \leq k \leq n} |S_k| \geq \varepsilon \right\} = \sum_{k=1}^n \mathbf{P}(E_k)$$

According to (5) of Sec. 26,

$$\mathbf{D}S_n = \sum_{k=0}^n \mathbf{P}(E_k) \cdot \mathbf{M}(S_n^2/E_k) \geq \sum_{k=1}^n \mathbf{P}(E_k) \cdot \mathbf{M}(S_n^2/E_k)$$

Clearly, then,

$$\begin{aligned} \mathbf{M}(S_n^2/E_k) &= \mathbf{M} \left\{ S_k^2 + 2 \sum_{j>k} S_k \eta_j + \sum_{j>k} \eta_j^2 + 2 \sum_{j>h>k} \eta_j \eta_h / E_k \right\} \geq \\ &\geq \mathbf{M} \left\{ S_k^2 + 2 \sum_{j>k} S_k \eta_j + 2 \sum_{j>h>k} \eta_j \eta_h / E_k \right\} \end{aligned}$$

Since the occurrence of the event E_k imposes a restriction solely on the values of the first k of the variables ξ_i , while the subsequent variables remain (given this condition) independent of one another and of S_k , it follows that

$$\mathbf{M}(S_k \eta_j / E_k) = \mathbf{M}(S_k / E_k) \cdot \mathbf{M}(\eta_j / E_k) = 0$$

and

$$\mathbf{M}(\eta_j \eta_h / E_k) = 0 \quad (h \neq j, \quad h > k, \quad j > k \geq 1)$$

Also, in accordance with (10), the inequality

$$\mathbf{M}(S_k^2 / E_k) \geq \varepsilon^2 \quad (k \geq 1)$$

holds.

We can therefore write

$$\mathbf{D}S_n \geq \varepsilon^2 \sum_{k=1}^n \mathbf{P}\{E_k\}$$

Whence

$$\sum_{k=1}^n \mathbf{P}\{E_k\} = \mathbf{P}\left\{\max_{1 \leq k \leq n} |S_k| \geq \varepsilon\right\} \leq \frac{1}{\varepsilon^2} \mathbf{D}S_n$$

This completes the proof of Kolmogorov's inequality.

Kolmogorov's Theorem. *If a sequence of mutually independent random variables ξ_1, ξ_2, \dots satisfies the condition*

$$\sum_{n=1}^{\infty} \frac{\mathbf{D}\xi_n}{n^2} < +\infty$$

then it obeys the strong law of large numbers.

Proof. We set

$$S_n = \sum_{k=1}^n (\xi_k - \mathbf{M}\xi_k), \quad v_n = \frac{1}{n} S_n$$

Consider the probability

$$P_m = \mathbf{P}\{\max |v_n| \leq \varepsilon, \quad 2^m \leq n < 2^{m+1}\}$$

Since

$$P_m \leq \mathbf{P}\{\max |S_n| \leq 2^m \varepsilon, \quad 2^m \leq n < 2^{m+1}\}$$

we have, by Kolmogorov's inequality,

$$P_m \leq \frac{1}{(2^m \varepsilon)^2} \sum_{j < 2^{m+1}} \mathbf{D}\xi_j$$

In accord with the remark made with respect to the lemma of this section, to prove the theorem it is sufficient for us to find that the following series converges:

$$\sum_{m=1}^{\infty} P_m$$

But according to the foregoing,

$$\sum_{m=1}^{\infty} P_m \leq \sum_{m=1}^{\infty} \frac{1}{(2^m \varepsilon)^2} \sum_{j < 2^{m+1}} D\xi_j = \frac{1}{\varepsilon^2} \sum_{j=1}^{\infty} D\xi_j \sum_j 2^{-2m}$$

where the sum \sum_j is extended over those values of m for which $2^{m+1} > j$.

We determine the number ρ by means of the inequalities

$$2^{\rho} \leq j < 2^{\rho+1}$$

Then

$$\sum_j 2^{-2m} = \sum_{m=\rho}^{\infty} 2^{-2m} = \frac{4}{3} \cdot 2^{-2\rho} < \frac{16}{3j^2}$$

Thus,

$$\sum_{m=1}^{\infty} P_m \leq \frac{16}{3\varepsilon^2} \sum_{j=1}^{\infty} \frac{D\xi_j}{j^2}$$

The proof of the theorem is consequently complete.

The theorem just proved clearly contains the following result:

Corollary. *If the variances of the random variables ξ_k are bounded by one and the same constant C , the sequence of mutually independent random variables $\xi_1, \xi_2, \xi_3, \dots$ obeys the strong law of large numbers.*

We thus see that the strong law of large numbers holds not only for the Bernoulli scheme with constant probability of occurrence of an event A in each of the trials (Borel's theorem) but also in the case of the Poisson scheme (the probability of event A depends on the number of the trial).

The theorem just proved enables us to obtain as a corollary one final result, which was also found by A. N. Kolmogorov.

Theorem. *The existence of expectation is a necessary and sufficient condition for applying the strong law of large numbers to a sequence of identically distributed and mutually independent random variables.*

Proof. From the existence of expectation follows the finiteness of the integral $\int |x| dF(x)$, where $F(x)$ is the distribution function of the random variables ξ_n . Therefore

$$\begin{aligned} \sum_{n=1}^{\infty} \mathbf{P} \{ |\xi| > n \} &= \sum_{n=1}^{\infty} \sum_{k \geq n} \mathbf{P} \{ k < |\xi| \leq k+1 \} = \\ &= \sum_{n=1}^{\infty} k \mathbf{P} \{ k < |\xi| \leq k+1 \} \leq \sum_{k=0}^{\infty} k \int_{k < |x| \leq k+1} |x| dF(x) < \\ &< \int |x| dF(x) < \infty \end{aligned} \quad (11)$$

We introduce the random variables

$$\xi_n^* = \begin{cases} \xi_n & \text{for } |\xi_n| \leq n \\ 0 & \text{for } |\xi_n| > n \end{cases}$$

We then obtain

$$D\xi_n^* \leq M\xi_n^{*2} = \int_{-n}^{+n} x^2 dF(x) \leq \sum_{k=0}^n (k+1)^2 \mathbf{P}\{k < |\xi| \leq k+1\}$$

and

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{D\xi_n^*}{n^2} &\leq \sum_{n=1}^{\infty} \sum_{k=0}^n \frac{(k+1)^2}{n^2} \mathbf{P}\{k < |\xi| \leq k+1\} \leq \\ &\leq \sum_{k=0}^{\infty} \mathbf{P}\{k < |\xi| \leq k+1\} (k+1)^2 \sum_{n \geq k} \frac{1}{n^2} \end{aligned}$$

Since

$$\sum_{n \geq k} \frac{1}{n^2} < \frac{1}{k^2} + \frac{1}{k} < \frac{2}{k}$$

we find, by (11),

$$\sum_{n=1}^{\infty} \frac{D\xi_n^*}{n^2} < \infty$$

That is, ξ_n^* satisfies the strong law of large numbers.

It remains to show that this proves the theorem. To do this, it will obviously be sufficient to detect that the probability of at least one inequality

$$\xi_n \neq \xi_n^*$$

for $n \geq N$ tends to zero as $N \rightarrow \infty$. Indeed,

$$\begin{aligned} \mathbf{P}\{\xi_n \neq \xi_n^* \text{ for any } n \geq N\} &\leq \sum_{n \geq N} \mathbf{P}\{\xi_n \neq \xi_n^*\} = \\ &= \sum_{n \geq N} \mathbf{P}\{|\xi_n| > n\} \leq \sum_{n=N}^{\infty} (n-N+1) \mathbf{P}\{n \leq |\xi_n| < n+1\} \leq \\ &\leq \sum_{n=N}^{\infty} n \int_{n \leq |x| < n+1} dF(x) \leq \sum_{n=N}^{\infty} \int_{n \leq |x| < n+1} |x| dF(x) = \int_{|x| \geq N} |x| dF(x) \end{aligned}$$

By hypothesis, the right-hand side of this inequality may be made smaller than any preassigned number by choosing N sufficiently large.

The fundamental role of the strong law of large numbers in the theory of probability and in its applications is exceedingly great. Indeed, suppose for a moment that, say, in the case of identically distributed summands having a finite expectation, the strong law does not hold. Then we can assert with probability arbitrarily close to unity that instances will recur where the arithmetic mean of the observational

results will be far removed from the expectation. And this would happen even in cases when the observations are performed without any systematic error and with complete definiteness. Would it then be possible to consider that the arithmetic mean of the observational results is close to the quantity being measured? And could we take it that under these conditions the arithmetic mean might be considered as an approximate value of the quantity being measured? This is doubtful.

EXERCISES

1. Prove that if the random variable ξ is such that $M e^{a\xi}$ exists ($a > 0$ is a constant), then

$$P \{ \xi \geq \varepsilon \} \leq \frac{M e^{a\varepsilon}}{e^{ae}}$$

2. Let $f(x) > 0$ be a nondecreasing function. Prove that if $M(f(|\xi - M\xi|))$ exists, then

$$P \{ |\xi - M\xi| \geq \varepsilon \} \leq \frac{M f(|\xi - M\xi|)}{f(\varepsilon)}$$

3. A sequence of independent and identically distributed random variables $\{\xi_i\}$ is defined by the equalities

$$(a) P \{ \xi_n = 2^{k - \log k - 2 \log \log k} \} = \frac{1}{2^k} \quad (k = 1, 2, 3, \dots)$$

$$(b) P \{ \xi_n = k \} = \frac{c}{k^2 \log^2 k} \left(k \geq 2, c^{-1} = \sum_{k=2}^{\infty} \frac{1}{k^2 \log^2 k} \right)$$

Prove that the law of large numbers is applicable to both sequences.

4. Prove that the law of large numbers may be applied to a sequence of independent random variables $\{\xi_n\}$ such that

$$P \{ \xi_n = n^\alpha \} = P \{ \xi_n = -n^\alpha \} = \frac{1}{2}$$

if and only if $\alpha < \frac{1}{2}$.

5. Prove that if the independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ are such that

$$\max_{1 \leq k \leq n} \int_{|x| \geq A} |x| dF_k(x) \rightarrow 0 \text{ when } A \rightarrow \infty$$

then the law of large numbers is applicable to the sequence $\{\xi_n\}$.

Hint. Employ the method used in the proof of Khinchin's theorem.

6. Using the result of the preceding problem, prove that if for a sequence of independent random variables $\{\xi_n\}$ there exist numbers $\alpha > 1$ and β such that $M|\xi|^\alpha \leq \beta$, then the law of large numbers is applicable to the sequence $\{\xi_n\}$ (*Markov's theorem*).

7. Given a sequence of random variables ξ_1, ξ_2, \dots , for which $D\xi_n \leq C$ and $R_{ij} \rightarrow 0$ as $|i - j| \rightarrow \infty$ (R_{ij} is the correlation coefficient of ξ_i and ξ_j), prove that the law of large numbers is applicable to this sequence (*Bernstein's theorem*).

Characteristic Functions

We have seen in the preceding chapters that probability theory makes wide use of the methods and the analytical apparatus of various divisions of mathematical analysis. A simple solution of an extremely wide range of problems of probability theory, especially those associated with the summation of independent random variables, is obtainable by means of *characteristic functions*, the theory of which has been developed in analysis and is known by the name of *Fourier transformations*. This chapter deals with the basic properties of characteristic functions.

Sec. 35. Definition and Elementary Properties of Characteristic Functions

The characteristic function of a random variable ξ is defined as the expectation of the random variable $e^{it\xi}$. If $F(x)$ is the distribution function of the variable ξ , the characteristic function is, by the theorem of Sec. 27,

$$f(t) = \int e^{itx} dF(x) \quad (1)$$

We agree, henceforth, to denote the characteristic function and the corresponding distribution function by the same letters, lower-case for the former and upper-case for the latter.

From the fact that $|e^{itx}|=1$ for all real t there follows the existence of the integral (1) for all distribution functions; hence, a characteristic function may be defined for every random variable.

* The letter t stands for a real parameter. The expectation of the complex random variables $\xi + i\eta$ is defined as $M\xi + iM\eta$. It is easy to verify that Theorems 1, 2, and 3 of Sec. 28 are valid in this case as well.

Theorem 1. *A characteristic function is uniformly continuous over the whole line and satisfies the following relations:*

$$f(0) = 1, \quad |f(t)| \leq 1 \quad (-\infty < t < \infty) \quad (2)$$

Proof. The relations (2) follow immediately from the definition of a characteristic function. Indeed, by (1)

$$f(0) = \int 1 \cdot dF(x) = 1$$

and

$$|f(t)| = \left| \int e^{itx} dF(x) \right| \leq \int |e^{itx}| dF(x) = \int dF(x) = 1$$

It now remains to prove the uniform continuity of the function $f(t)$. For this purpose let us consider the difference

$$f(t+h) - f(t) = \int e^{itx} (e^{ixh} - 1) dF(x)$$

and let us estimate its absolute value. We have

$$|f(t+h) - f(t)| \leq \int |e^{ixh} - 1| dF(x)$$

Let $\varepsilon > 0$ be arbitrary; we choose A sufficiently large so that

$$\int_{|x| > A} dF(x) < \frac{\varepsilon}{4}$$

and select h so small that for $|x| < A$

$$|e^{ixh} - 1| < \frac{\varepsilon}{2}$$

Then

$$|f(t+h) - f(t)| \leq \int_{-A}^A |e^{ixh} - 1| dF(x) + 2 \int_{|x| > A} dF(x) \leq \varepsilon$$

This inequality proves the theorem.

Theorem 2. *If $\eta = a\xi + b$, where a and b are constants, then*

$$f_{\eta}(t) = f_{\xi}(at) e^{ibt}$$

where $f_{\eta}(t)$ and $f_{\xi}(t)$ denote the characteristic functions of the variables η and ξ .

Proof. Indeed,

$$f_{\eta}(t) = \mathbf{M}e^{it\eta} = \mathbf{M}e^{it(a\xi+b)} = e^{itb} \mathbf{M}e^{ita\xi} = e^{itb} f_{\xi}(at)$$

Theorem 3. *The characteristic function of the sum of two independent random variables is equal to the product of their characteristic functions.*

Proof. Let ξ and η be independent random variables and let $\zeta = \xi + \eta$. Then, clearly, $e^{it\xi}$ and $e^{it\eta}$ will also be random variables along with ξ and η . From this it follows that

$$\mathbf{M}e^{it\zeta} = \mathbf{M}e^{it(\xi+\eta)} = \mathbf{M}e^{it\xi}e^{it\eta} = \mathbf{M}e^{it\xi}\mathbf{M}e^{it\eta}$$

This proves the theorem.

Corollary. *If*

$$\xi = \xi_1 + \xi_2 + \dots + \xi_n$$

and each term is independent of the sum of the preceding terms, then the characteristic function of the variable ξ is equal to the product of the characteristic functions of the summands.

The application of characteristic functions rests to a large extent on the property formulated in Theorem 3. As we saw in Sec. 24, the addition of independent random variables leads to an extremely complicated operation, the convolution of the distribution functions of the summands. With regard to characteristic functions, this complex operation is replaced by an extremely simple one, the simple multiplication of characteristic functions.

Theorem 4. *If a random variable ξ has an absolute moment of the n th order, then the characteristic function of the variable ξ is differentiable n times and when $k \leq n$*

$$f^{(k)}(0) = i^k \mathbf{M}\xi^k \quad (3)$$

Proof. Indeed, formal differentiation of the characteristic function k times ($k \leq n$) leads to the equation

$$f^{(k)}(t) = i^k \int x^k e^{itx} dF(x) \quad (4)$$

But

$$\left| \int x^k e^{itx} dF(x) \right| \leq \int |x|^k dF(x)$$

and, consequently, by hypothesis of the theorem, it is bounded. It follows from this that the integral (4) exists and differentiation is legitimate. Putting $t=0$ in (4) we find

$$f^{(k)}(0) = i^k \int x^k dF(x)$$

Expectation and variance are very simply expressed by means of derivatives of the logarithm of the characteristic function. Indeed, put

$$\psi(t) = \log f(t)$$

Then

$$\psi'(t) = \frac{f'(t)}{f(t)}$$

and

$$\psi''(t) = \frac{f''(t) \cdot f(t) - [f'(t)]^2}{f^2(t)}$$

Taking into account Equation (3) and that $f(0) = 1$, we find

$$\psi'(0) = f'(0) = iM\xi$$

and

$$\psi''(0) = f''(0) - [f'(0)]^2 = i^2 M\xi^2 - [iM\xi]^2 = -D\xi$$

Whence

$$\left. \begin{aligned} M\xi &= \frac{1}{i} \psi'(0) \\ D\xi &= -\psi''(0) \end{aligned} \right\} \quad (5)$$

The k th derivative of the logarithm of the characteristic function at the point 0, multiplied by i^k , is called the *cumulant (semi-invariant) of the k th order of the random variable*.

As follows directly from Theorem 3, when independent random variables are added their cumulants are added too.

We have just seen that the first two cumulants are the expectation and the variance, that is, the first-order moment and a certain rational function of the moments of first and second orders. It will be readily seen, by means of computation, that the cumulant of any order k is an entire rational function of the first k moments. By way of illustration, we give the explicit expressions of cumulants of the third and fourth orders:

$$i^3 \psi'''(0) = -\{M\xi^3 - 3M\xi^2 \cdot M\xi + 2[M\xi]^3\}$$

$$i^4 \psi^{IV}(0) = M\xi^4 - 4M\xi^3 M\xi - 3[M\xi^2]^2 + 12M\xi^2 [M\xi]^2 - 6[M\xi]^4$$

We now consider a few examples of characteristic functions.

Example 1. A random variable ξ is distributed in accordance with the normal law with expectation a and variance σ^2 . The characteristic function of the variable ξ is

$$\varphi(t) = \frac{1}{\sigma \sqrt{2\pi}} \int e^{itx - \frac{(x-a)^2}{2\sigma^2}} dx$$

By the substitution

$$z = \frac{x-a}{\sigma} - it\sigma$$

$\varphi(t)$ is reduced to the form

$$\varphi(t) = e^{iat - \frac{\sigma^2 t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty - it\sigma}^{\infty - it\sigma} e^{-\frac{z^2}{2}} dz$$

It is known that for any real α

$$\int_{-\infty - i\alpha}^{\infty - i\alpha} e^{-\frac{z^2}{2}} dz = \sqrt{2\pi}$$

hence,

$$\varphi(t) = e^{iat - \frac{\sigma^2 t^2}{2}}$$

Using Theorem 4 we can readily compute the central moments for a normal distribution and in this alternative way obtain the result of the example considered in Sec. 30.

Example 2. Find the characteristic function of a random variable ξ that is Poisson distributed.

By hypothesis, the variable ξ assumes only integral values, and

$$P\{\xi = k\} = \frac{\lambda^k e^{-\lambda}}{k!} \quad (k = 0, 1, 2, \dots)$$

where $\lambda > 0$ is a constant.

The characteristic function of the variable ξ is

$$\begin{aligned} f(t) = M e^{it\xi} &= \sum_{k=0}^{\infty} e^{ikt} P\{\xi = k\} = \sum_{k=0}^{\infty} e^{ikt} \frac{\lambda^k}{k!} e^{-\lambda} = \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda + \lambda e^{it}} = e^{\lambda(e^{it} - 1)} \end{aligned}$$

According to (5) we then find

$$M\xi = \frac{1}{i} \psi'(0) = \lambda, \quad D\xi = -\psi''(0) = \lambda$$

The first of these equations was earlier obtained by us directly (see Example 3, Sec. 26).

Example 3. A random variable ξ is uniformly distributed over the interval $(-a, a)$. The characteristic function is equal to

$$f(t) = \int_{-a}^a e^{itx} \frac{dx}{2a} = \frac{\sin at}{at}$$

Example 4. Find the characteristic function of the variable μ , which is equal to the number of occurrences of an event A in n independent trials, in each of which the probability that A will occur is p .

The variable μ may be represented in the form of a sum:

$$\mu = \mu_1 + \mu_2 + \dots + \mu_n$$

of n independent variables, each of which takes on only two values, 0 and 1, with respective probabilities $q=1-p$ and p . The variable μ_k assumes the value 1 if event A occurs in the k th trial and the value 0 if event A does not occur in the k th trial.

The characteristic function of μ_k is equal to

$$f_k(t) = \mathbf{M}e^{it\mu_k} = e^{it \cdot 0}q + e^{it \cdot 1}p = q + pe^{it}$$

According to Theorem 3, the characteristic function of the variable μ is

$$f(t) = \prod_{k=1}^n f_k(t) = (q + pe^{it})^n$$

Let us also find the characteristic function of the variable $\eta = \frac{\mu - np}{\sqrt{npq}}$. By Theorem 2 it is

$$\begin{aligned} f_\eta(t) &= e^{-it} \sqrt{\frac{np}{q}} f\left(\frac{t}{\sqrt{npq}}\right) = e^{-it} \sqrt{\frac{np}{q}} \left(q + pe^{it \frac{t}{\sqrt{npq}}}\right)^n = \\ &= \left(qe^{-it} \sqrt{\frac{p}{nq}} + pe^{it} \sqrt{\frac{q}{np}}\right)^n \end{aligned}$$

Example 5. Characteristic functions satisfy the equation $f(-t) = \overline{f(t)}$.

Indeed,

$$f(-t) = \int e^{-itx} dF(x) = \overline{\int e^{itx} dF(x)} = \overline{f(t)}$$

Sec. 36. The Inversion Formula and the Uniqueness Theorem

We have seen that from the distribution function of a random variable ξ it is always possible to find its characteristic function. For us it is important that the converse proposition holds as well: a distribution function is uniquely determined by its characteristic function.

Theorem 1. Let $f(t)$ and $F(x)$ be the characteristic function and the distribution function, respectively, of a random variable ξ . If x_1 and x_2 are points of continuity of the function $F(x)$, then

$$F(x_2) - F(x_1) = \frac{1}{2\pi} \lim_{c \rightarrow \infty} \int_{-c}^c \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t) dt \quad (1)$$

Proof. From the definition of a characteristic function it follows that the integral

$$J_c = \frac{1}{2\pi} \int_{-c}^c \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t) dt$$

is equal to

$$J_c = \frac{1}{2\pi} \int_{-c}^c \int \frac{1}{it} [e^{it(z-x_1)} - e^{it(z-x_2)}] dF(z) dt$$

The order of integration may be changed in the last integral, since the integral converges absolutely with respect to z , and the limits of integration are finite with respect to t . Thus

$$\begin{aligned} J_c &= \frac{1}{2\pi} \int \left[\int_{-c}^c \frac{e^{it(z-x_1)} - e^{it(z-x_2)}}{it} dt \right] dF(z) = \\ &= \frac{1}{2\pi} \int \left[\int_0^c \frac{e^{it(z-x_1)} - e^{-it(z-x_1)} - e^{it(z-x_2)} + e^{-it(z-x_2)}}{it} dt \right] dF(z) = \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \int_0^c \left[\frac{\sin t(z-x_1)}{t} - \frac{\sin t(z-x_2)}{t} \right] dt dF(z) \end{aligned}$$

From analysis it is a well-known fact that as $c \rightarrow \infty$

$$\frac{1}{\pi} \int_0^c \frac{\sin \alpha t}{t} dt \rightarrow \begin{cases} \frac{1}{2} & \text{if } \alpha > 0 \\ -\frac{1}{2} & \text{if } \alpha < 0 \end{cases} \quad (2)$$

and this convergence is uniform with respect to α in each region $\alpha > \delta > 0$ (or $\alpha < -\delta$) and when $|\alpha| \leq \delta$, for all c ,

$$\left| \frac{1}{\pi} \int_0^c \frac{\sin \alpha t}{t} dt \right| < 1 \quad (3)$$

Assume for the sake of definiteness that $x_2 > x_1$ and represent the integral J_c as the following sum:

$$J_c = \int_{-\infty}^{x_1-\delta} + \int_{x_1-\delta}^{x_1+\delta} + \int_{x_1+\delta}^{x_2-\delta} + \int_{x_2-\delta}^{x_2+\delta} + \int_{x_2+\delta}^{\infty} \psi(c, z; x_1, x_2) dF(z)$$

where for brevity we have used the notation

$$\psi(c, z; x_1, x_2) = \frac{1}{\pi} \int_0^c \left\{ \frac{\sin t(z-x_1)}{t} - \frac{\sin t(z-x_2)}{t} \right\} dt$$

and $\delta > 0$ is chosen so that $x_1 + \delta < x_2 - \delta$.

The inequalities $z - x_1 < -\delta$ and $z - x_2 < -\delta$ hold in the region $-\infty < z < x_1 - \delta$. We therefore conclude, on the basis of (2), that

as $c \rightarrow \infty$

$$\int_{-\infty}^{x_1-\delta} \psi(c, z; x_1, x_2) dF(z) \rightarrow 0$$

Similarly, when $x_2 + \delta < z < +\infty$ and when $c \rightarrow \infty$,

$$\int_{x_2+\delta}^{\infty} \psi(c, z; x_1, x_2) dF(z) \rightarrow 0$$

Further, since in the region $x_1 + \delta < z < x_2 - \delta$ the inequalities $z - x_1 > \delta$ and $z - x_2 < \delta$ are valid, it follows from (2) that as $c \rightarrow \infty$,

$$\int_{x_1+\delta}^{x_2-\delta} \psi(c, z; x_1, x_2) dF(z) \rightarrow \int_{x_1+\delta}^{x_2-\delta} dF(z) = F(x_2 - \delta) - F(x_1 + \delta)$$

Finally, by (3) we can take advantage of the estimates

$$\left| \int_{x_1-\delta}^{x_1+\delta} \psi(c, z; x_1, x_2) dF(z) \right| < 2 \int_{x_1-\delta}^{x_1+\delta} dF(z) = 2[F(x_1 + \delta) - F(x_1 - \delta)]$$

and

$$\left| \int_{x_2-\delta}^{x_2+\delta} \psi(c, z; x_1, x_2) dF(z) \right| < 2 \int_{x_2-\delta}^{x_2+\delta} dF(z) = 2[F(x_2 + \delta) - F(x_2 - \delta)]$$

We thus find that for every $\delta > 0$

$$\overline{\lim}_{c \rightarrow \infty} J_c = F(x_2 - \delta) - F(x_1 + \delta) + R_1(\delta, x_1, x_2)$$

and

$$\underline{\lim}_{c \rightarrow \infty} J_c = F(x_2 - \delta) - F(x_1 + \delta) + R_2(\delta, x_1, x_2)$$

where

$$|R_i(\delta, x_1, x_2)| < 2\{F(x_1 + \delta) - F(x_1 - \delta) + F(x_2 + \delta) - F(x_2 - \delta)\} \\ (i = 1, 2)$$

Now let $\delta \rightarrow 0$. From the fact that x_1 and x_2 are points of continuity of the function $F(x)$ there follow the equations

$$\lim_{\delta \rightarrow 0} F(x_1 + \delta) = \lim_{\delta \rightarrow 0} F(x_1 - \delta) = F(x_1)$$

and

$$\lim_{\delta \rightarrow 0} F(x_2 + \delta) = \lim_{\delta \rightarrow 0} F(x_2 - \delta) = F(x_2)$$

And since J_c does not depend on δ , it follows that

$$\lim_{c \rightarrow \infty} J_c = F(x_2) - F(x_1)$$

Equation (1) is called the *inversion formula*. We shall use it to derive the following important proposition (the *uniqueness theorem*).

Theorem 2. *A distribution function is uniquely determined by its characteristic function.*

Proof. Indeed, it follows directly from Theorem 1 that the following formula holds at each point of continuity of the function $F(x)$:

$$F(x) = \frac{1}{2\pi} \lim_{y \rightarrow -\infty} \lim_{c \rightarrow \infty} \int_{-c}^{+c} \frac{e^{-ity} - e^{-itx}}{it} f(t) dt$$

where the limit in y is evaluated with respect to any set of points y that are points of continuity of the function $F(x)$.

As an application of the last theorem we will prove the following propositions.

Example 1. If the independent random variables ξ_1 and ξ_2 are normally distributed, then their sum $\xi = \xi_1 + \xi_2$ is also distributed normally.

Indeed, if

$$M\xi_1 = a_1, D\xi_1 = \sigma_1^2; M\xi_2 = a_2, D\xi_2 = \sigma_2^2$$

then the characteristic functions of the variables ξ_1 and ξ_2 are

$$f_1(t) = e^{ia_1t - \frac{1}{2}\sigma_1^2t^2}, \quad f_2(t) = e^{ia_2t - \frac{1}{2}\sigma_2^2t^2}$$

By Theorem 3, Sec. 34, the characteristic function $f(t)$ of the sum is equal to

$$f(t) = f_1(t) \cdot f_2(t) = e^{it(a_1+a_2) - \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}$$

This is the characteristic function of a normal law with expectation $a = a_1 + a_2$ and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$. On the basis of the uniqueness theorem we conclude that the distribution function of the variable ξ is normal.

The converse proposition, due to H. Cramér, that we formulated in Sec. 24 may be stated as follows in terms of characteristic functions: *if $f_1(t)$ and $f_2(t)$ are characteristic functions and*

$$f_1(t) \cdot f_2(t) = e^{-\frac{t^2}{2}}$$

then

$$f_1(t) = e^{iat - \sigma^2 \frac{t^2}{2}}, \quad f_2(t) = e^{-iat - \frac{(1-\sigma^2)t^2}{2}} \quad (0 \leq \sigma \leq 1)$$

Example 2. The independent random variables ξ_1 and ξ_2 are distributed according to the Poisson law, and

$$\mathbf{P}\{\xi_1 = k\} = \frac{\lambda_1^k e^{-\lambda_1}}{k!}, \quad \mathbf{P}\{\xi_2 = k\} = \frac{\lambda_2^k e^{-\lambda_2}}{k!}$$

Prove that the random variable $\xi = \xi_1 + \xi_2$ is distributed in accordance with the Poisson law with parameter $\lambda = \lambda_1 + \lambda_2$.

Indeed, in Example 2 of the preceding section we found that the characteristic functions of the random variables ξ_1 and ξ_2 are

$$f_1(t) = e^{\lambda_1(e^{it} - 1)}, \quad f_2(t) = e^{\lambda_2(e^{it} - 1)}$$

By Theorem 3 of the preceding section, the characteristic function of the sum $\xi = \xi_1 + \xi_2$ is

$$f(t) = f_1(t) \cdot f_2(t) = e^{(\lambda_1 + \lambda_2)(e^{it} - 1)}$$

that is, it is a characteristic function of some Poisson law. By the uniqueness theorem, the only distribution with $f(t)$ as its characteristic function is the Poisson law for which

$$\mathbf{P}\{\xi = k\} = \frac{(\lambda_1 + \lambda_2)^k e^{-(\lambda_1 + \lambda_2)}}{k!} \quad (k \geq 0)$$

D. A. Raikov proved the more profound converse proposition: *if the sum of two independent random variables is distributed according to the Poisson law, then each summand is also distributed in accordance with the Poisson law.*

Example 3. A characteristic function is real when and only when the corresponding distribution function is symmetric, that is, when the distribution function satisfies the equation

$$F(x) = 1 - F(-x + 0)$$

for all x .

If the distribution function is symmetric, then its characteristic function is real. If ξ has a symmetric distribution function, then both ξ and $-\xi$ are identically distributed. Hence the equation

$$f(t) = \mathbf{M}e^{it\xi} = \mathbf{M}e^{-it\xi} = f(-t) = \overline{f(t)}$$

holds and this means that $f(t)$ is real.

To prove the converse proposition, consider the random variable $\eta = -\xi$. The distribution function of the variable η is

$$G(x) = P\{\eta < x\} = \mathbf{P}\{\xi > -x\} = 1 - F(-x + 0)$$

The characteristic functions of the variables ξ and η are connected by the relation

$$g(t) = Me^{it\eta} = Me^{-it\xi} = \overline{Me^{it\xi}} = \overline{f}(t)$$

Since by hypothesis $f(t)$ is real, it follows that $\overline{f}(t) = f(t)$ and, hence,

$$g(t) = f(t)$$

From the uniqueness theorem we now conclude that the distribution functions of the variables ξ and η coincide, that is, that

$$F(x) = 1 - F(-x + 0)$$

and the proof is complete.

Theorem 3. *If a characteristic function $|f(t)|$ is integrable over the entire line, then the corresponding distribution function $F(x)$ is absolutely continuous, its derivative $p(x)$ is continuous and*

$$p(x) = F'(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt$$

Proof. If the function $f(t)$ is summable over the entire line, then the function $\frac{e^{-itx_1} - e^{-itx_2}}{it} f(t)$ has this property as well, and for this reason the inversion formula may be written as

$$F(x_2) - F(x_1) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-itx_1} - e^{-itx_2}}{it} f(t) dt$$

Now let h be such that $x_1 = x - h$ and $x_2 = x + h$ are continuity points of $F(x)$. After simple formal transformations we arrive at the equation

$$F(x+h) - F(x-h) = 2h \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin th}{th} e^{-itx} f(t) dt \quad (4)$$

Since $\left| \frac{\sin th}{th} \right| \leq 1$, it follows that

$$F(x+h) - F(x-h) \leq 2h \cdot \frac{1}{2\pi} \int_{-\infty}^{\infty} |f(t)| dt$$

This last inequality obviously proves that $F(x)$ is absolutely continuous.

Now (4) may be expressed as follows:

$$\frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin th}{th} e^{-itx} f(t) dt \quad (5)$$

Since the integrand converges to $e^{-itx}f(t)$ as $h \rightarrow 0$, it follows from the well-known Lebesgue theorem, on passing to the limit under the integral sign, that

$$\lim_{h \rightarrow 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\sin th}{th} e^{-itx} f(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt$$

Since the limit of the right-hand side of Equation (5) exists, the limit of its left-hand side also exists. Thus, for every value of x ,

$$p(x) = \lim_{h \rightarrow \infty} \frac{F(x+h) - F(x-h)}{2h} = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} f(t) dt$$

By straightforward computation it follows that

$$|p(x+h) - p(x)| \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} 2 \left| \sin \frac{th}{2} \right| |f(t)| dt$$

Evaluate the integral on the right-hand side. To do that write it in the form of a sum:

$$\frac{1}{\pi} \int_{|t| < A} \left| \sin \frac{th}{2} \right| |f(t)| dt + \frac{1}{\pi} \int_{|t| > A} \left| \sin \frac{th}{2} \right| |f(t)| dt$$

Let $\epsilon > 0$ be given. Choose A so large that

$$\frac{1}{\pi} \int_{|t| > A} |f(t)| dt < \frac{\epsilon}{2}$$

The first integral may be made less than $\epsilon/2$ by choosing h sufficiently small. This completes the proof of the theorem.

Sec. 37. Helly's Theorems

In the sequel we shall require two theorems of a purely analytic nature: the first and second Helly theorems.

Let us agree that a sequence of nondecreasing functions

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

converges weakly to a nondecreasing function $F(x)$ if as $n \rightarrow \infty$ it converges to this function at every one of its points of continuity.

Henceforth we will always assume that the functions $F_n(x)$ satisfy the supplementary condition

$$F_n(-\infty) = 0$$

and will not mention this fact each time.

We straightway note that for weak convergence it is sufficient that the sequence of functions converge to the function $F(x)$ on some everywhere-dense set D . Indeed, let x be any point and x' and x'' be some two points of the set D such that $x' \leq x \leq x''$. Also that

$$F_n(x') \leq F_n(x) \leq F_n(x'')$$

Consequently,

$$\lim_{n \rightarrow \infty} F_n(x') \leq \lim_{n \rightarrow \infty} F_n(x) \leq \overline{\lim}_{n \rightarrow \infty} F_n(x) \leq \lim_{n \rightarrow \infty} F_n(x'')$$

And since by hypothesis

$$\lim_{n \rightarrow \infty} F_n(x') = F(x') \quad \text{and} \quad \lim_{n \rightarrow \infty} F_n(x'') = F(x'')$$

it also follows that

$$F(x') \leq \lim_{n \rightarrow \infty} F_n(x) \leq \overline{\lim}_{n \rightarrow \infty} F_n(x) \leq F(x'')$$

But the middle terms in these inequalities do not depend on x' and x'' , and so

$$F(x-0) \leq \lim_{n \rightarrow \infty} F_n(x) \leq \overline{\lim}_{n \rightarrow \infty} F_n(x) \leq F(x+0)$$

If the function $F(x)$ is continuous at the point x , then

$$F(x-0) = F(x) = F(x+0)$$

Consequently, at continuity points of the function $F(x)$,

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

Helly's First Theorem. *Any sequence of uniformly bounded non-decreasing functions*

$$F_1(x), F_2(x), \dots, F_n(x), \dots \quad (1)$$

contains at least one subsequence

$$F_{n_1}(x), F_{n_2}(x), \dots, F_{n_k}(x), \dots$$

that converges weakly to some nondecreasing function $F(x)$.

Proof. Let D be some countable everywhere-dense set of points $x'_1, x'_2, \dots, x'_n, \dots$. Take the values of the functions of the sequence (1) at the point x'_1 :

$$F_1(x'_1), F_2(x'_1), \dots, F_n(x'_1), \dots$$

Since, by hypothesis, the set of these values is bounded, it contains at least one subsequence

$$F_{11}(x'_1), F_{12}(x'_1), \dots, F_{1n}(x'_1), \dots \quad (2)$$

that converges to a certain limiting value, which we denote by $G(x'_1)$. We now consider the set of numbers

$$F_{11}(x'_2), F_{12}(x'_2), \dots, F_{1n}(x'_2), \dots$$

Since this set is bounded as well, there exists in it a subsequence that converges to some limiting value $G(x'_2)$. Thus we can extract from the sequence (2) a subsequence

$$F_{21}(x), F_{22}(x), \dots, F_{2n}(x), \dots \quad (3)$$

for which simultaneously $\lim_{n \rightarrow \infty} F_{2n}(x'_1) = G(x'_1)$ and $\lim_{n \rightarrow \infty} F_{2n}(x'_2) = G(x'_2)$.

We continue this extraction of subsequences

$$F_{k1}(x), F_{k2}(x), \dots, F_{kn}(x), \dots \quad (4)$$

for which the equations $\lim_{n \rightarrow \infty} F_{kn}(x'_r) = G(x'_r)$ hold simultaneously for all $r \leq k$. Now construct the diagonal sequence

$$F_{11}(x), F_{22}(x), \dots, F_{nn}(x), \dots \quad (5)$$

The whole of it has ultimately been extracted from the sequence (1), and so for it $\lim_{n \rightarrow \infty} F_{nn}(x'_1) = G(x'_1)$. Further, since the entire diagonal sequence, with the exception of the first term only, has been extracted from the sequence (2), it follows that $\lim_{n \rightarrow \infty} F_{nn}(x'_2) = G(x'_2)$. Generally speaking, the entire diagonal sequence, with the exception of the first $k-1$ terms, has been extracted from the sequence (4), and so for it, too, the $\lim_{n \rightarrow \infty} F_{nn}(x'_k) = G(x'_k)$ holds for every k . The result may be formulated thus: the sequence (1) contains at least one subsequence which converges at all points x'_k of the set D to some function $G(x)$ defined on D . And since the functions $F_{nn}(x)$ do not decrease and are uniformly bounded, it is obvious that the function $G(x)$ as well will be nondecreasing and bounded. It is now clear that the function $G(x)$ defined on the set D may be continued so that it will be defined over the entire line $-\infty < x < \infty$, while remaining nondecreasing and bounded.

The sequence (5) converges to this function on the everywhere-dense set D ; hence, it converges weakly to it, which is what we set out to prove. It will be noted that the function obtained by continuing the function G may prove not to be continuous from the left. But we can change its values at the discontinuity points so as to restore this property. The subsequence F_{nn} will converge weakly to the thus "corrected" function.

Helly's Second Theorem. *Let $f(x)$ be a continuous function and let the sequence of nondecreasing uniformly bounded functions*

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

converge weakly to the function $F(x)$ on some finite interval $a \leq x \leq b$, where a and b are continuity points of the function $F(x)$; then

$$\lim_{n \rightarrow \infty} \int_a^b f(x) dF_n(x) = \int_a^b f(x) dF(x)$$

Proof. From the continuity of the function $f(x)$ it follows that for any positive constant ε there will be a subdivision of the interval $a \leq x \leq b$ by the points $x_0 = a, x_1, \dots, x_N = b$ into subintervals $x_k \leq x \leq x_{k+1}$ such that in each interval (x_k, x_{k+1}) the inequality $|f(x) - f(x_k)| < \varepsilon$ will hold. Taking advantage of this circumstance, we can introduce an auxiliary function $f_*(x)$ that takes on only a finite number of values and define it by means of the equalities

$$f_*(x) = f(x_k) \text{ for } x_k \leq x < x_{k+1}$$

Clearly, for all x in the interval $a \leq x \leq b$ the inequality

$$|f(x) - f_*(x)| < \varepsilon$$

holds. In doing so we can select beforehand the subdivision points x_1, x_2, \dots, x_{N-1} so that they will be continuity points of the function $F(x)$. By virtue of the convergence of the functions $F_1(x), F_2(x), F_3(x), \dots$ to the function $F(x)$, the following inequalities will hold at all subdivision points for n sufficiently large:

$$|F(x_k) - F_n(x_k)| < \frac{\varepsilon}{MN} \quad (6)$$

where M is the maximum of the absolute value of $f(x)$ in the interval $a \leq x \leq b$.

It is clear without explanation that

$$\begin{aligned} \left| \int_a^b f(x) dF(x) - \int_a^b f(x) dF_n(x) \right| &\leq \\ &\leq \left| \int_a^b f(x) dF(x) - \int_a^b f_*(x) dF(x) \right| + \\ &+ \left| \int_a^b f_*(x) dF(x) - \int_a^b f_*(x) dF_n(x) \right| + \\ &+ \left| \int_a^b f_*(x) dF_n(x) - \int_a^b f(x) dF_n(x) \right| \end{aligned}$$

It is easy to compute that the first summand on the right-hand side does not exceed $\varepsilon [F(b) - F(a)]$ and the third does not exceed

$\varepsilon [F_n(b) - F_n(a)]$. But the second summand is found to be equal to

$$\left| \sum_{k=0}^{N-1} f(x_k) [F(x_{k+1}) - F(x_k)] - \sum_{k=0}^{N-1} f(x_k) [F_n(x_{k+1}) - F_n(x_k)] \right| =$$

$$= \left| \sum_{k=0}^{N-1} f(x_k) [F(x_{k+1}) - F_n(x_{k+1})] - \sum_{k=0}^{N-1} f(x_k) [F(x_k) - F_n(x_k)] \right|$$

and, consequently, for n sufficiently large it does not exceed 2ε , as follows from the inequality (6). By virtue of the uniform boundedness of the function $F_n(x)$, the sum

$$\varepsilon [F(b) - F(a)] + \varepsilon [F_n(b) - F_n(a)] + 2\varepsilon$$

can be made arbitrarily small together with ε .

The Generalized Second Theorem of Helly. *If the function $f(x)$ is continuous and bounded over the entire line $-\infty < x < \infty$, the sequence of uniformly bounded nondecreasing functions*

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

converges weakly to the function $F(x)$ and

$$\lim_{n \rightarrow \infty} F_n(-\infty) = F(-\infty), \quad \lim_{n \rightarrow \infty} F_n(+\infty) = F(+\infty)$$

it follows that

$$\lim_{n \rightarrow \infty} \int f(x) dF_n(x) = \int f(x) dF(x)$$

Proof. Let $A < 0$ and $B > 0$; we put

$$J_1 = \left| \int_{-\infty}^A f(x) dF(x) - \int_{-\infty}^A f(x) dF_n(x) \right|$$

$$J_2 = \left| \int_A^B f(x) dF(x) - \int_A^B f(x) dF_n(x) \right|$$

$$J_3 = \left| \int_B^{\infty} f(x) dF(x) - \int_B^{\infty} f(x) dF_n(x) \right|$$

It is obvious that

$$\left| \int f(x) dF(x) - \int f(x) dF_n(x) \right| \leq J_1 + J_2 + J_3$$

The quantities J_1 and J_3 may be made arbitrarily small if one chooses A and B sufficiently large in absolute value and also such that the points A and B are points of continuity of the function $F(x)$, and by choosing n sufficiently large. Indeed, let M be the

upper bound of $|f(x)|$ for $-\infty < x < \infty$; then

$$J_1 \leq M [F(A) + F_n(A)]$$

$$J_3 \leq M [F(+\infty) - F(B)] + M [F_n(+\infty) - F_n(B)]$$

But

$$\lim_{A \rightarrow -\infty} F(A) = 0, \quad \lim_{B \rightarrow \infty} F(B) = F(+\infty)$$

And since by hypothesis

$$\lim_{n \rightarrow \infty} F_n(A) = F(A), \quad \lim_{n \rightarrow \infty} F_n(B) = F(B)$$

our assertion about J_1 and J_3 is proved. For n sufficiently large, the quantity J_2 may be made arbitrarily small by virtue of Helly's theorem for a finite interval.

The theorem is proved.

Sec. 38. Limit Theorems for Characteristic Functions

From the point of view of the applications of characteristic functions to the derivation of asymptotic formulas of probability theory, two limit theorems (direct and converse) are of prime importance. These theorems state that the correspondence existing between distribution functions and characteristic functions is not only one-to-one but also continuous.

The Direct Limit Theorem. *If a sequence of distribution functions*

$$F_1(x), F_2(x), \dots, F_n(x), \dots$$

converges weakly to the distribution function $F(x)$, then the sequence of characteristic functions

$$f_1(t), f_2(t), \dots, f_n(t), \dots$$

converges to the characteristic function $f(t)$. This convergence is uniform in each finite interval of t .

Proof. Since

$$f_n(t) = \int e^{itx} dF_n(x), \quad f(t) = \int e^{itx} dF(x)$$

and the function e^{itx} is continuous and bounded over the entire line $-\infty < t < \infty$, according to the generalized second theorem of Helly, as $n \rightarrow \infty$,

$$f_n(t) \rightarrow f(t)$$

The assertion that this convergence is uniform in every finite interval of t is verified literally by the same arguments used in proving Helly's second theorem.

The Converse Limit Theorem. *If a sequence of characteristic functions*

$$f_1(t), f_2(t), \dots, f_n(t), \dots \quad (1)$$

converges to the continuous function $f(t)$, then the sequence of distribution functions

$$F_1(x), F_2(x), \dots, F_n(x), \dots \quad (2)$$

converges weakly to some distribution function $F(x)$ [by virtue of the direct limit theorem $f(t) = \int e^{itx} dF(x)$].

Proof. On the basis of Helly's first theorem we conclude that the sequence (2) definitely contains a subsequence

$$F_{n_1}(x), F_{n_2}(x), \dots, F_{n_k}(x), \dots \quad (3)$$

which converges weakly to some nondecreasing function $F(x)$. It is clear in this case that the function $F(x)$ may be considered continuous on the left:

$$\lim_{x' \rightarrow x-0} F(x') = F(x)$$

Generally speaking, the function $F(x)$ need not be a distribution function, since for this to be the case the conditions $F(-\infty)=0$ and $F(+\infty)=1$ must also hold. Indeed, for the sequence of functions

$$F_n(x) = \begin{cases} 0 & \text{for } x \leq -n \\ \frac{1}{2} & \text{for } -n < x \leq n \\ 1 & \text{for } x > n \end{cases}$$

the limit function is $F(x) \equiv \frac{1}{2}$, and, consequently, $F(-\infty)$ and $F(+\infty)$ are also equal to $\frac{1}{2}$. However, as will now be shown, under the conditions of our theorem we will definitely have $F(-\infty)=0$ and $F(+\infty)=1$.

Indeed, if this were not so, then, taking into account that for the limit function $F(x)$ the relations $F(-\infty) \geq 0$ and $F(+\infty) \leq 1$ must hold, we would have

$$\delta = F(+\infty) - F(-\infty) < 1$$

Now take some positive number ε less than $1-\delta$. Since by hypothesis the sequence of characteristic functions (1) converges to the function $f(t)$ it follows that $f(0)=1$. And since also the function $f(t)$ is continuous, one can choose a positive number τ so small

that the inequality

$$\frac{1}{2\tau} \left| \int_{-\tau}^{\tau} f(t) dt \right| > 1 - \frac{\varepsilon}{2} > \delta + \frac{\varepsilon}{2} \quad (4)$$

will hold. But at the same time we can choose $X > \frac{4}{\tau\varepsilon}$ and K so large that for $k > K$

$$\delta_k = F_{n_k}(X) - F_{n_k}(-X) < \delta + \frac{\varepsilon}{4}$$

Since $f_{n_k}(t)$ is a characteristic function, it follows that

$$\int_{-\tau}^{\tau} f_{n_k}(t) dt = \int \left[\int_{-\tau}^{\tau} e^{itx} dt \right] dF_{n_k}(x)$$

The integral on the right-hand side of this equation may be evaluated as follows. On the one hand, since $|e^{itx}| = 1$,

$$\left| \int_{-\tau}^{\tau} e^{itx} dt \right| \leq 2\tau$$

On the other hand,

$$\int_{-\tau}^{\tau} e^{itx} dt = \frac{2}{x} \sin \tau x$$

and since $|\sin \tau x| \leq 1$, for $|x| > X$

$$\left| \int_{-\tau}^{\tau} e^{itx} dt \right| < \frac{2}{X}$$

From this, using the first estimate for $|x| \leq X$ and the second for $|x| > X$, we get

$$\begin{aligned} \left| \int_{-\tau}^{\tau} f_{n_k}(t) dt \right| &\leq \left| \int_{|x| \leq X} \left(\int_{-\tau}^{\tau} e^{itx} dt \right) dF_{n_k}(x) \right| + \\ &\quad + \left| \int_{|x| > X} \left(\int_{-\tau}^{\tau} e^{itx} dt \right) dF_{n_k}(x) \right| < 2\tau\delta_k + \frac{2}{X} \end{aligned}$$

and, hence,

$$\frac{1}{2\tau} \left| \int_{-\tau}^{\tau} f_{n_k}(t) dt \right| < \delta + \frac{\varepsilon}{2}$$

This inequality continues to hold in the limit as well:

$$\frac{1}{2\tau} \left| \int_{-\tau}^{\tau} f(t) dt \right| \leq \delta + \frac{\varepsilon}{2}$$

which obviously contradicts inequality (4).

Thus, the function $F(x)$, to which the sequence $F_{n_k}(x)$ weakly converges, is a distribution function; by the direct limit theorem its characteristic function is $f(t)$.

To complete the proof of the theorem it remains to prove that the entire sequence (2) also converges weakly to the function $F(x)$. We assume that this is not so. Then there will be a subsequence of functions

$$F_{n'_1}(x), F_{n'_2}(x), \dots, F_{n'_k}(x), \dots \quad (5)$$

that converges weakly to some function $F^*(x)$ different from $F(x)$ in at least one of its points of continuity. From what has already been proved $F^*(x)$ must be a distribution function with the characteristic function $f(t)$. By the uniqueness theorem we should have

$$F^*(x) = F(x)$$

This contradicts the assumption.

We note that the hypothesis of the theorem is satisfied in each of the following two cases:

(1) The sequence of characteristic functions $f_n(t)$ converges to some function $f(t)$ uniformly in every finite interval of t .

(2) The sequence of characteristic functions $f_n(t)$ converges to a characteristic function $f(t)$.

Example. To illustrate the use of limit theorems we consider the proof of the integral theorem of DeMoivre-Laplace.

In Example 4, Sec. 35, we found the characteristic function of the random variable $\eta = \frac{\mu - np}{\sqrt{npq}}$:

$$f_n(t) = \left(qe^{-it} \sqrt{\frac{p}{nq}} + pe^{it} \sqrt{\frac{q}{np}} \right)^n$$

Taking advantage of the expansion in a Maclaurin series, we find

$$qe^{-it} \sqrt{\frac{p}{nq}} + pe^{it} \sqrt{\frac{q}{np}} = 1 - \frac{t^2}{2n} (1 + R_n)$$

where

$$R_n = 2 \sum_{k=3}^{\infty} \frac{1}{k!} \left(\frac{it}{\sqrt{n}} \right)^{k-2} \frac{pq^k + q(-p)^k}{\sqrt{(pq)^k}}$$

Since $R_n \rightarrow 0$ as $n \rightarrow \infty$, it follows that

$$f_n(t) = \left[1 - \frac{t^2}{2n} (1 + R_n) \right]^n \rightarrow e^{-\frac{t^2}{2}}$$

By virtue of the *converse limit theorem*, it follows from the foregoing that for every x

$$\mathbf{P} \left\{ \frac{\mu - np}{\sqrt{npq}} < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

as $n \rightarrow \infty$.

From the continuity of the limit function it is easy to deduce that this convergence will be uniform in x .

Sec. 39. Positive Definite Functions.

The purpose of this section is to give an exhaustive description of the class of characteristic functions. The basic theorem given below was discovered by A. Ya. Khinchin and S. Bochner at the same time and was first published by S. Bochner.

To formulate and prove this theorem we have to introduce a new concept. We will say that the continuous function $f(t)$ of the real argument t is *positively defined* in the interval $-\infty < t < \infty$ if, for any real numbers t_1, t_2, \dots, t_n , complex numbers $\xi_1, \xi_2, \dots, \xi_n$ and integer n ,

$$\sum_{k=1}^n \sum_{j=1}^n f(t_j - t_k) \xi_j \bar{\xi}_k \geq 0 \quad (1)$$

We list a few of the most elementary properties of positive definite functions.

1. $f(0) \geq 0$. Indeed, put $n=1$, $t_1=0$, $\xi_1=1$; then from the condition of positive definiteness of the function $f(t)$ we find

$$\sum_{k=1}^n \sum_{j=1}^n f(t_k - t_j) \xi_k \bar{\xi}_j = f(0) \geq 0$$

2. For any real t ,

$$f(-t) = \overline{f(t)}$$

To prove this, in (1) put $n=2$, $t_1=0$, $t_2=t$, and ξ_1, ξ_2 arbitrary. By hypothesis we have

$$\begin{aligned} 0 &\leq \sum_{k=1}^2 \sum_{j=1}^2 f(t_k - t_j) \xi_k \bar{\xi}_j = \\ &= f(0-0) \xi_1 \bar{\xi}_1 + f(0-t) \xi_1 \bar{\xi}_2 + f(t-0) \xi_2 \bar{\xi}_1 + f(t-t) \xi_2 \bar{\xi}_2 = \\ &= f(0) (|\xi_1|^2 + |\xi_2|^2) + f(-t) \xi_1 \bar{\xi}_2 + f(t) \bar{\xi}_1 \xi_2 \quad (2) \end{aligned}$$

and so the quantity

$$f(-t)\xi_1\bar{\xi}_2 + f(t)\bar{\xi}_1\xi_2$$

must be real. Thus, if we put $f(-t) = \alpha_1 + i\beta_1$, $f(t) = \alpha_2 + i\beta_2$, $\xi_1\bar{\xi}_2 = \gamma + i\delta$, $\bar{\xi}_1\xi_2 = \gamma - i\delta$, then it must be that

$$\alpha_1\delta + \beta_1\gamma - \alpha_2\delta + \beta_2\gamma = 0$$

Since ξ_1 and ξ_2 and, hence, γ and δ are arbitrary, it must be that $\alpha_1 - \alpha_2 = 0$ and $\beta_1 + \beta_2 = 0$.

From this our assertion follows.

3. For all real t

$$|f(t)| \leq f(0)$$

In inequality (2) put $\xi_1 = f(t)$, $\xi_2 = -|f(t)|$; then from the preceding result

$$2f(0)|f(t)|^2 - |f(t)|^2|f(t)| - |f(t)|^2|f(t)| \geq 0$$

From this we get, when $|f(t)| \neq 0$,

$$f(0) \geq |f(t)|$$

But if $|f(t)| = 0$, then again by virtue of Property 1, we have

$$f(0) \geq |f(t)|.$$

From what has been proved it follows incidentally that if a positive definite function is such that $f(0) = 0$, then $f(t) \equiv 0$.

Bochner-Khinchin Theorem. *For a continuous function $f(t)$ satisfying the condition $f(0) = 1$ to be characteristic, it is necessary and sufficient that it be positive definite.*

Proof. In one direction the theorem is trivial. Indeed, if

$$f(t) = \int e^{ixt} dF(x)$$

where $F(x)$ is some distribution function, then for any integral n , arbitrary real t_1, t_2, \dots, t_n , and complex numbers $\xi_1, \xi_2, \dots, \xi_n$ we have

$$\begin{aligned} \sum_{k=1}^n \sum_{j=1}^n f(t_k - t_j) \xi_k \bar{\xi}_j &= \sum_{k=1}^n \sum_{j=1}^n \left\{ \int e^{ix(t_k - t_j)} dF(x) \right\} \xi_k \bar{\xi}_j = \\ &= \int \sum_{k=1}^n \sum_{j=1}^n e^{ix(t_k - t_j)} \xi_k \bar{\xi}_j dF(x) = \\ &= \int \left(\sum_{k=1}^n e^{it_k x} \xi_k \right) \left(\sum_{j=1}^n e^{-it_j x} \bar{\xi}_j \right) dF(x) = \int \left| \sum_{k=1}^n e^{it_k x} \xi_k \right|^2 dF(x) \geq 0 \end{aligned}$$

The sufficiency proof requires a more involved reasoning.

The proof given here is taken from Yu. V. Linnik's book. To a considerable extent it relies on Theorem 3, Sec. 36, the limit theorems for characteristic functions, and on the following lemma.

Lemma. *If the function $\varphi(t)$ is measurable, bounded and summable on the interval $(-T, T)$ and*

$$f(x) = \int_{-T}^T e^{-itx} \varphi(t) dt \geq 0 \quad (3)$$

then the function $f(x)$ is integrable on the entire line.

Proof. The function $f(x)$ being continuous, it is integrable over any finite interval. Put

$$G(x) = \int_{-x}^x f(z) dz$$

By virtue of the nonnegativity of $f(z)$ the function $G(z)$ is non-decreasing, and so to prove the lemma it suffices to demonstrate that $G(x)$ is bounded. For this purpose consider the function

$$F(u) = \frac{1}{u} \int_u^{2u} G(x) dx$$

Clearly,

$$F(u) \geq \frac{G(u)}{u} \int_u^{2u} dx = G(u)$$

and, consequently, if it is shown that the function $F(x)$ is bounded, then that will prove the lemma.

It is easy to verify that

$$G(x) = 2 \int_{-T}^T \frac{\sin xt}{t} \varphi(t) dt$$

and

$$F(x) = \frac{4}{x} \int_{-T}^T \frac{\sin^2 xt}{t^2} \varphi(t) dt - \frac{4}{x} \int_{-T}^T \frac{\sin^2 \frac{xt}{2}}{t^2} \varphi(t) dt$$

Let $M = \sup |\varphi(x)|$; then

$$\frac{4}{x} \int_{-T}^T \frac{\sin^2 xt}{t^2} \varphi(t) dt \leq 4M \int_{-\infty}^{\infty} \frac{\sin^2 u}{u^2} du$$

It is obvious that

$$\frac{4}{x} \int_{-T}^T \frac{\sin^2 \frac{xt}{2}}{t^2} \varphi(t) dt \leq 2M \int_{-\infty}^{\infty} \frac{\sin^2 u}{u^2} du$$

The boundedness of the functions $F(x)$ and $G(x)$ is proved, and this proves the lemma.

Now assume that the function $f(t)$ is positive definite, continuous and $f(0) = 1$. Let $Z > 0$. We consider the function

$$p_Z(x) = \frac{1}{2\pi Z} \int_0^Z \int_0^Z f(u-v) e^{-iux} e^{-ivx} du dv$$

The function $p_Z(x)$ must be nonnegative by virtue of the positive definiteness of $f(t)$ (the double integral is the limit of the corresponding sums). If in the double integral we make the change of variables

$$t = u - v, \quad z = u$$

and perform elementary transformations, it will be found that

$$p_Z(x) = \frac{1}{2\pi} \int_{-Z}^Z \left(1 - \frac{|t|}{Z}\right) f(t) e^{-itx} dt$$

Since the function $p_Z(x)$ is nonnegative and representable as (3), the lemma just proved may be applied to it. And so $p_Z(x)$ is integrable over the entire line. The function $f(t)$ is continuous and therefore from Theorem 3, Sec. 36, it follows that

$$\left(1 - \frac{|t|}{Z}\right) f(t) = \int_{-\infty}^{\infty} p_Z(x) e^{itx} dx$$

for all t ($|t| \leq Z$). In particular, for $t = 0$

$$\int_{-\infty}^{\infty} p_Z(x) dx = f(0) = 1$$

Thus, $p_Z(x)$ is the density of some probability distribution and so, $\left(1 - \frac{|t|}{Z}\right) f(t)$ is the corresponding characteristic function. For $Z \rightarrow \infty$, the functions

$$\left(1 - \frac{|t|}{Z}\right) f(t)$$

uniformly converge to the function $f(t)$ in every finite interval of t . From this it follows that $f(t)$ is a characteristic function.

This completes the proof of the theorem.

Sec. 40. Characteristic Functions of Multidimensional Random Variables

In this section we give, without proof, basic information concerning the characteristic functions of multidimensional random variables.

The *characteristic function* of an n -dimensional random variable $(\xi_1, \xi_2, \dots, \xi_n)$ is defined as the expectation of the variable $e^{i(t_1\xi_1+t_2\xi_2+\dots+t_n\xi_n)}$, where t_1, t_2, \dots, t_n are real variables:

$$f(t_1, t_2, \dots, t_n) = \mathbf{M} \exp \left(i \sum_{k=1}^n t_k \xi_k \right) \quad (1)$$

If $F(x_1, x_2, \dots, x_n)$ is the distribution function of the variable $(\xi_1, \xi_2, \dots, \xi_n)$, then, as we know from the preceding result*,

$$f(t_1, t_2, \dots, t_n) = \int \dots \int \left(\exp i \sum_{k=1}^n t_k x_k \right) dF(x_1, \dots, x_n) \quad (2)$$

As in the one-dimensional case, the characteristic function of an n -dimensional random variable is uniformly continuous over the entire space $(-\infty < t_j < +\infty, 1 \leq j \leq n)$ and satisfies the following relations:

$$\begin{aligned} f(0, 0, \dots, 0) &= 1 \\ |f(t_1, t_2, \dots, t_n)| &\leq 1 \quad (-\infty < t_k < +\infty, k = 1, 2, \dots) \\ f(-t_1, -t_2, \dots, -t_n) &= \overline{f(t_1, t_2, \dots, t_n)} \end{aligned}$$

From the characteristic function $f(t_1, t_2, \dots, t_n)$ of the random variable $(\xi_1, \xi_2, \dots, \xi_n)$ it is easy to find the characteristic function of any k -dimensional ($k < n$) variable $(\xi_{j_1}, \xi_{j_2}, \dots, \xi_{j_k})$ whose components are the variables ξ_s ($1 \leq s \leq n$). To do this, in formula (2) one has to put equal to zero all arguments t_s for $s \neq j_r$ ($1 \leq r \leq k$). Thus, for example, the characteristic function of the variable ξ_1 is

$$f_1(t_1) = f(t_1, 0, \dots, 0)$$

It follows from the definition that if the components of the variable $(\xi_1, \xi_2, \dots, \xi_n)$ are *independent* random variables, then its characteristic function is equal to the product of the characteristic functions of the components

$$f(t_1, t_2, \dots, t_n) = f_1(t_1) f_2(t_2) \dots f_n(t_n)$$

* Compare Theorem 1, Sec. 27, and the remark on multidimensional Stieltjes integrals in Sec. 26.

Just as in the one-dimensional case, multidimensional characteristic functions make it easy to find moments of various orders. Thus, for example,

$$\begin{aligned} M_{\xi_1 \xi_2 \dots \xi_n}^{k_1 k_2 \dots k_n} &= \int \int \dots \int x_1^{k_1} x_2^{k_2} \dots x_n^{k_n} dF(x_1, x_2, \dots, x_n) = \\ &= (i) \sum_1^n k_j \left[\frac{\partial^{k_1+k_2+\dots+k_n} f(t_1, t_2, \dots, t_n)}{\partial t_1^{k_1} \partial t_2^{k_2} \dots \partial t_n^{k_n}} \right]_{t_1=t_2=\dots=t_n=0} \end{aligned}$$

For the computation of characteristic functions it is useful to know the following theorem which the reader will be able to prove without any difficulty.

Theorem 1. *If the characteristic function of a variable $(\xi_1, \xi_2, \dots, \xi_n)$ is equal to $f(t_1, t_2, \dots, t_n)$, then the characteristic function of the variable $(\sigma_1 \xi_1 + a_1, \sigma_2 \xi_2 + a_2, \dots, \sigma_n \xi_n + a_n)$, where a_i and $\sigma_i (1 \leq i \leq n)$ are real constants, is*

$$\exp \left(i \sum_{k=1}^n a_k t_k \right) \cdot f(\sigma_1 t_1, \sigma_2 t_2, \dots, \sigma_n t_n)$$

Example 1. Let us calculate the characteristic function of a two-dimensional random variable distributed according to the normal law:

$$p(x, y) = \frac{1}{2\pi \sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} [x^2 - 2rxy + y^2] \right\} \quad (3)$$

By formula (2),

$$f(t_1, t_2) = \int \int e^{i(t_1 x + t_2 y)} p(x, y) dx dy$$

Changing variables we can reduce $f(t_1, t_2)$ to the form

$$f(t_1, t_2) = e^{-\frac{1}{2}(t_1^2 + 2rt_1 t_2 + t_2^2)} \frac{1}{2\pi} \int \int e^{-\frac{1}{2}(u^2 + v^2)} du dv = e^{-\frac{1}{2}(t_1^2 + 2rt_1 t_2 + t_2^2)}$$

Example 2. Applying Theorem 1, we find the characteristic function of the variable (η_1, η_2) which is distributed according to the normal law:

$$\begin{aligned} p(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \\ &\times \exp \left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x-a)^2}{\sigma_1^2} - 2r \frac{(x-a)(y-b)}{\sigma_1\sigma_2} + \frac{(y-b)^2}{\sigma_2^2} \right] \right\} \quad (4) \end{aligned}$$

If we put $\eta_1 = \sigma_1 \xi_1 + a$, $\eta_2 = \sigma_2 \xi_2 + b$, then the variable (ξ_1, ξ_2) will be distributed according to the law (3). According to Theorem 1,

the characteristic function of the variable (η_1, η_2) is

$$\varphi(t_1, t_2) = \exp \left[iat_1 + iat_2 - \frac{1}{2} (\sigma_1^2 t_1^2 + 2\sigma_1 \sigma_2 r t_1 t_2 + \sigma_2^2 t_2^2) \right]$$

The following theorem is a consequence of the definition of a characteristic function.

Theorem 2. *If $f(t_1, t_2, \dots, t_n)$ is the characteristic function of the variable $(\xi_1, \xi_2, \dots, \xi_n)$, then the characteristic function of the sum $\xi_1 + \xi_2 + \dots + \xi_n$ is*

$$f(t) = f(t, t, \dots, t)$$

Note. We notice that

$$f(t) = f(tt_1, tt_2, \dots, tt_n)$$

is the characteristic function of the sum $t_1 \xi_1 + t_2 \xi_2 + \dots + t_n \xi_n$.

Example 3. We apply Theorem 2 to determine the distribution of the sum $\eta_1 + \eta_2$ if (η_1, η_2) is distributed in accordance with the law (4).

According to Theorem 2 the characteristic function of the sum $\eta_1 + \eta_2$ is

$$f(t) = \exp \left[it(a+b) - \frac{t^2}{2} (\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2) \right]$$

We know from Example 1 of Sec. 35 that this is the characteristic function of the normal law with expectation $a+b$ and variance $\sigma_1^2 + 2r\sigma_1\sigma_2 + \sigma_2^2$. Earlier, we obtained this result directly (Example 2, Sec. 24).

At the beginning of this chapter we saw that the characteristic function of a sum of independent random variables is equal to the product of the characteristic functions of the summands. We shall show that this property is only a necessary but not a sufficient condition for the independence of random variables. For this purpose consider the two-dimensional random variable (ξ, η) , whose density function may be expressed as

$$p(x, y) = p_1(x) p_2(y) + \varphi(x) \psi(y) - \varphi(y) \psi(x)$$

where $p_1(x)$ and $p_2(y)$ are one-dimensional density functions, and $\varphi(x)$ and $\psi(x)$ ($\varphi(x) \neq \psi(x)$) are odd integrable functions. It is easy to see that such density functions exist. Indeed, the function

$$p(x, y) = \frac{1}{4} e^{-|x|-|y|} \{ 1 + xye^{-2|x|-2|y|} - xye^{-|x|-2|y|} \}$$

is just such an instance. It satisfies the inequality $p(x, y) > 0$ for all x and y and, moreover,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y) dx dy = 1$$

The random variables ξ and η are dependent since their joint density function cannot be expressed as the product of two factors, each of which depends only on one argument x or y . The density function of the component ξ is

$$p_{\xi}(x) = \int_{-\infty}^{\infty} p(x, y) dy = p_1(x)$$

and the density function of the component η is

$$p_{\eta}(y) = \int_{-\infty}^{\infty} p(x, y) dx = p_2(y)$$

The two-dimensional characteristic function for the vector (ξ, η) is equal to

$$f(t, \tau) = f_1(t) f_2(\tau) + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{itx + i\tau y} \varphi(x) \psi(y) dy dx - \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{itx + i\tau y} \psi(x) \varphi(y) dx dy$$

where

$$f_1(t) = \int_{-\infty}^{\infty} e^{itx} p_1(x) dx, \quad f_2(\tau) = \int_{-\infty}^{\infty} e^{i\tau y} p_2(y) dy$$

In our particular example, the characteristic function of the vector is

$$f(t, \tau) = \frac{1}{(1+t^2)(1+\tau^2)} + 24t\tau \left[\frac{1}{(4+t^2)(9+\tau^2)} - \frac{1}{(9+t^2)(4+\tau^2)} \right]$$

The characteristic function of the sum $\xi + \eta$ is, according to Theorem 2, equal in the general case to

$$f(t, t) = f_1(t) f_2(t)$$

that is to say, it is equal to the product of the characteristic functions of the summands. We have thus shown that there exist independent random variables for which the characteristic function of the sum is equal to the product of the characteristic functions of the summands.

It is important to note that in the multidimensional case the following theorem holds.

Theorem 3. *A distribution function $F(x_1, x_2, \dots, x_n)$ is uniquely determined by its characteristic function.*

The proof of this proposition is based on the inversion formula.

Theorem 4. *If $f(t_1, t_2, \dots, t_n)$ is the characteristic function and $F(x_1, x_2, \dots, x_n)$ is the distribution function of the random variable*

$(\xi_1, \xi_2, \dots, \xi_n)$, then

$$\begin{aligned} & \mathbf{P} \{a_k \leq \xi_k < b_k, \quad k=1, 2, \dots, n\} = \\ & = \lim_{T \rightarrow \infty} \frac{1}{(2\pi)^n} \int_{-T}^T \int_{-T}^T \dots \int_{-T}^T \prod_{k=1}^n \frac{e^{it_k a_k} - e^{it_k b_k}}{it_k} f(t_1, \dots, t_n) dt_1 dt_2 \dots dt_n \end{aligned}$$

where a_k and b_k are any real numbers that satisfy only one requirement: that the probability of falling on the surface of a parallelepiped $a_k \leq \xi_k < b_k$ ($k=1, 2, \dots, n$) be equal to zero.

Just as in the one-dimensional case, we have the direct and the converse limit theorems for the characteristic functions. We shall not dwell on this.

Example 4. One says that an n -dimensional random variable $(\xi, \xi_1, \dots, \xi_n)$ has a *nondegenerate* (proper) *n -dimensional normal distribution* if its density function is of the form

$$p(x_1, x_2, \dots, x_n) = C e^{-\frac{1}{2} Q(x_1, x_2, \dots, x_n)}$$

where

$$Q(x_1, x_2, \dots, x_n) = \sum_{i,j} b_{ij} (x_i - a_i)(x_j - a_j)$$

is a positive definite quadratic form, and C , a_i and b_{ij} are real constants.

Simple computations demonstrate* that

$$C = (\sqrt{2\pi})^{-n} \sqrt{D}$$

where

$$D = \begin{vmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{vmatrix}$$

Denote by D_{ij} the minor D , which corresponds to the element b_{ij} , then

$$\begin{aligned} \mathbf{M}\xi_j &= a_j, \quad \sigma_j^2 = \mathbf{D}\xi_j = \frac{D_{jj}}{D} \quad (j=1, 2, \dots, n) \\ r_{ij} &= \frac{\mathbf{M}(\xi_i - a_i)(\xi_j - a_j)}{\sigma_i \sigma_j} = \frac{D_{ij}}{\sqrt{D_{ii} D_{jj}}} \quad (i, j=1, 2, \dots, n) \end{aligned}$$

The determinant D and its principal minors are positive.

* The usual procedure for such computations is to change variables, which reduces the form Q to a sum of squares, and to carry out all the computations in the new variables.

Using ordinary computations it is easy to verify that the characteristic function of the variable $(\xi_1, \xi_2, \dots, \xi_n)$ is equal to

$$f(t_1, t_2, \dots, t_n) = e^{i \sum_{j=1}^n a_j t_j - \frac{1}{2} \sum_{k=1}^n \sum_{j=1}^n \sigma_j \sigma_k r_{jk} t_j t_k}$$

Thus, an n -dimensional normal distribution is completely determined by specifying the expectation and variance.

From the expression for the characteristic function of an n -dimensional normally distributed random variable we see that the distribution of the variable

$$(\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_k})$$

will, for all $1 \leq i_1 < i_2 < \dots < i_k \leq n$, be a k -dimensional normal distribution.

EXERCISES

1. Prove that the functions

$$f_1(t) = \sum_{k=0}^{\infty} a_k \cos kt, \quad f_2(t) = \sum_{k=0}^{\infty} a_k e^{i\lambda_k t}$$

where $a_k \geq 0$ and $\sum_{k=0}^{\infty} a_k = 1$ are characteristic functions; determine the corresponding probability distributions.

2. Find the characteristic function for the following probability density functions:

$$(a) \quad p(x) = \frac{a}{2} e^{-a|x|};$$

$$(b) \quad p(x) = \frac{a}{\pi(a^2 + x^2)};$$

$$(c) \quad p(x) = \begin{cases} 0 & \text{when } |x| \geq a \\ \frac{a - |x|}{a^2} & \text{when } |x| \leq a; \end{cases}$$

$$(d) \quad p(x) = \frac{2 \sin^2 \frac{ax}{2}}{\pi ax^2}$$

Note. The attentive reader will have noted that Examples (a) and (b) and also (c) and (d) are, so to speak, inverse.

3. Prove that the functions

$$\varphi_1(t) = \frac{1}{\cosh t}, \quad \varphi_2(t) = \frac{t}{\sinh t}, \quad \varphi_3(t) = \frac{1}{\cosh^2 t}$$

are characteristic functions of the density functions

$$\rho_1(x) = \frac{1}{2 \cosh \frac{\pi x}{2}}, \quad \rho_2(x) = \frac{\pi}{4 \cosh^2 \frac{\pi x}{2}}, \quad \rho_3(x) = \frac{x}{2 \sinh \frac{\pi x}{2}}$$

respectively.

4. Find the probability distributions of random variables whose characteristic functions are

$$(a) \cos t; (b) \cos^2 t; (c) \frac{a}{1+it}; (d) \frac{\sin at}{at}$$

5. Prove that the function defined by the equations

$$f(t) = f(-t), \quad f(t+2a) = f(t), \quad f(t) = \frac{a-t}{a} \text{ for } 0 \leq t \leq a$$

is a characteristic function.

Note. The characteristic functions of Examples 2 (d) and 5 possess the following remarkable property:

$$f_2(t) = f_5(t) \text{ for } |t| \leq a, \\ f_2(t) \neq f_5(t) \text{ for } |t| > a \text{ and } t \neq \pm 2a, \dots$$

There, thus, exist characteristic functions whose values coincide in an arbitrarily large interval $(-a, +a)$ and are not identically equal. The first instance of two such characteristic functions was pointed out by B. V. Gnedenko; then Krein indicated the necessary and sufficient conditions for which the identity of two characteristic functions follows from their equality in some interval $(-a, +a)$.

6. Prove that one can find independent random variables ξ_1, ξ_2, ξ_3 such that the probability distributions of ξ_2 and ξ_3 are different, while the distribution functions of the sums $\xi_1 + \xi_2$ and $\xi_1 + \xi_3$ are the same.

Hint. Make use of the results of Examples 2 (a) and 5.

7. Prove that if $f(t)$ is a characteristic function equal to zero when $|t| \geq a$, then the function $\varphi(t)$ defined by the equations

$$\varphi(t) = \begin{cases} f(t) & \text{when } |t| \leq a \\ f(t+2a) & \text{when } -\infty < t < \infty \end{cases}$$

is also a characteristic function.

Hint. Make use of the Bochner-Khinchin theorem.

8. Prove that if $f(t)$ is a characteristic function, then the function

$$\varphi(t) = e^{f(t)-1}$$

is also a characteristic function.

9. Prove that if the function $f(t)$ is a characteristic function, then the function

$$\varphi(t) = \frac{1}{t} \int_0^t f(z) dz$$

is also a characteristic function.

10. Prove that for any real characteristic function $\varphi(t)$ the inequality

$$1 - \varphi(2t) \leq 4 \{1 - \varphi(t)\}$$

holds and, hence, for any characteristic function the inequality

$$1 - |f(2t)|^2 \leq 4 \{1 - |f(t)|^2\}$$

also holds.

11. Prove that for any real characteristic function the inequality

$$1 + \varphi(2t) \geq 2 \{\varphi(t)\}^2$$

holds.

12. Prove that if $F(x)$ is a distribution and $f(t)$ the corresponding characteristic function, then for any value of x the equation

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(t) e^{-itx} dt = F(x+0) - F(x-0)$$

holds.

13. Prove that if $F(x)$ is a distribution function, $f(t)$ the corresponding characteristic function, and x_v are abscissas of jumps in the function $F(x)$, then

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |f(t)|^2 dt = \sum_v \{F(x_v+0) - F(x_v-0)\}$$

14. Prove that if a random variable has a density function, then its characteristic function tends to zero as $t \rightarrow \infty$.

15. A random variable (ξ) is Poisson distributed; $M\xi = \lambda$. Prove that as $\lambda \rightarrow \infty$, the distribution of the variable $\frac{\xi - \lambda}{\sqrt{\lambda}}$ tends to the normal law for which the parameters a and σ are $a=0$, $\sigma=1$.

16. A random variable ξ has the density function

$$p(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0 \end{cases}$$

Prove that as $\alpha \rightarrow \infty$ the distribution of the variable $\frac{\beta\xi - \alpha}{\sqrt{\alpha}}$ converges to

the normal distribution with parameters $a=0$, $\sigma=1$.

Note. The results of Exercises 15 and 16 permit using tables of normal distribution when computing the probabilities $P\{a \leq \xi < b\}$ for large values of λ (or α). In particular, it turns out that for a chi-square distribution the limiting relation gives excellent accuracy already for $n \geq 30$. This fact is constantly utilized in statistics.

17. Prove that if $\varphi(t)$ is a characteristic function and the function $\psi(t)$ is such that for some sequence $\{h_n\}$ ($h_n \rightarrow \infty$ as $n \rightarrow \infty$), the products

$$\varphi(t) \psi(h_n t) = f_n(t)$$

are also characteristic functions, then the function $\psi(t)$ is a characteristic function.

The Classical Limit Theorem

Sec. 41. Statement of the Problem

The integral limit theorem of DeMoivre-Laplace that we proved in Chapter 2 served as the source of a broad range of investigations of fundamental importance both to the theory of probability itself and to its multiplicity of applications in the natural sciences, technology and the economic sciences. To give an idea of the trend of these investigations, we shall restate the DeMoivre-Laplace theorem in a somewhat different form. Namely, if, as we have frequently done, we denote by μ_k the number of occurrences of an event A in the k th trial, then the number of occurrences of A in n successive trials is equal to $\sum_{k=1}^n \mu_k$. Further, in Example 5, Sec. 28, we computed that $M \sum_{k=1}^n \mu_k = np$ and $D \sum_{k=1}^n \mu_k = npq$. Therefore, the DeMoivre-Laplace theorem may be written as follows:

$$P \left\{ a \leq \frac{\sum_{k=1}^n (\mu_k - M\mu_k)}{\sqrt{\sum_{k=1}^n D\mu_k}} < b \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz \quad (1)$$

as $n \rightarrow \infty$. In words: *the probability that the sum of deviations of independent random variables—which take on two values, 0 and 1, with probabilities respectively equal to q and $p = 1 - q$ ($0 < p < 1$)—from their expectations divided by the square root of the sum of the variances of the summands will lie between the limits from a to*

b tends to the integral $\frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz$ uniformly in a and b as the number of summands increases to infinity.

The natural question arises: How closely tied up is the relation (1) with the special choice of summands μ_k ? Will it not hold in the case of weaker restrictions imposed on the distribution functions of the summands? The statement of this problem and also its solution belong in the main to P. L. Chebyshev and his pupils A. A. Markov and A. M. Lyapunov. Their investigations have shown that one should impose on the summands only the most general restrictions, the meaning of which depends on the fact that the separate summands should exert an insignificant effect on the sum. In the next section we will give a precise statement of this condition. The reasons why these results are so vastly important in applications lie in the very essence of mass-scale phenomena, the study of the regularities of which is, as we have already had occasion to say, the actual subject of the theory of probability.

One of the most important schemes used to exploit the results of probability theory in the natural sciences and technology consists in the following. It is assumed that a process occurs under the influence of a large number of independently operating random factors, each of which only to a negligible extent modifies the course of the phenomenon or process. The investigator who is interested in the process as a whole, and not the operation of separate factors, observes only the overall operation of these factors. We illustrate with two typical examples.

Example 1. Let a measurement be made. The result will unavoidably be influenced by a large number of factors that generate errors in the measurement. These will include errors due to the state of the measuring instrument, which might vary in gross fashion under the effect of various atmospheric or mechanical factors. There will also be human errors of the observer caused by peculiarities of vision or hearing and also those that might be altered slightly due to the psychic or physical state of the observer, and so forth. Each of these factors would generate a negligible error. But the measurement is affected at once by all these errors, the result being an "overall error". In other words, the actually observed error of measurement will be a random variable—the sum of an enormous number of negligibly small and independent random variables. And though these quantities are unknown, as also are their distribution functions, their effect on the results of the measurements is noticeable and for this reason must be the subject of study.

Example 2. In many industries large batches of identical articles are produced by the mass-production process. Let us consider some numerical characteristic of the product we are interested in. Insofar as the article conforms to certain technical standards, there is a certain standard value of the characteristic we have chosen. Actually, however, there is always observed a certain deviation from this stan-

dard value. In a properly organized production process, such deviations can only be caused by random factors, each of which produces only an unnoticeable effect. The overall action, however, generates a noticeable deviation from the norm.

Any number of such instances might be cited.

Thus, there arises the problem of studying regularities peculiar to sums of a large number of independent random variables, each of which exerts but a slight effect on the sum. Later on we will make the meaning of this requirement more precise. Instead of studying sums of a very large but finite number of summands, we will consider a sequence of sums with ever larger numbers of summands and assume that the solutions of the problems we are interested in are given by limiting distribution functions for a sequence of distribution functions of the sums. This kind of passage from a finite statement of the problem to a limiting statement is customary both in modern mathematics and in many divisions of the natural sciences.

We have thus arrived at a consideration of the following problem: given a sequence of mutually independent random variables

$$\xi_1, \xi_2, \dots, \xi_n, \dots$$

about which we suppose that they have finite expectations and variances. From now on we will adhere to the following notations:

$$a_k = M\xi_k, \quad b_k^2 = D\xi_k, \quad B_n^2 = \sum_{k=1}^n b_k^2 = D \sum_{k=1}^n \xi_k$$

The question is: what conditions must be imposed on the variables ξ_k so that the distribution functions of the sums

$$\frac{1}{B_n} \sum_{k=1}^n (\xi_k - a_k) \quad (2)$$

converge to the normal distribution law? In the next section we will see that for this purpose it is sufficient that the *Lindeberg condition* be satisfied: for any $\tau > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^2} \sum_{k=1}^n \int_{|x-a_k| > \tau B_n} (x-a_k)^2 dF_k(x) = 0$$

where $F_k(x)$ denotes the distribution function of the variable ξ_k .

Let us clarify the meaning of this condition.

Denote by A_k the event consisting in the fact that

$$|\xi_k - a_k| > \tau B_n \quad (k = 1, 2, \dots, n)$$

and estimate the probability

$$\mathbf{P} \left\{ \max_{1 \leq k \leq n} |\xi_k - a_k| > \tau B_n \right\}$$

Since

$$\mathbf{P} \left\{ \max_{1 \leq k \leq n} |\xi_k - a_k| > \tau B_n \right\} = \mathbf{P} \{ A_1 + A_2 + \dots + A_n \}$$

and

$$\mathbf{P} \{ A_1 + A_2 + \dots + A_n \} \leq \sum_{k=1}^n \mathbf{P} \{ A_k \}$$

by noting that

$$\mathbf{P} \{ A_k \} = \int_{|x-a_k| > \tau B_n} dF_k(x) \leq \frac{1}{(\tau B_n)^2} \int_{|x-a_k| > \tau B_n} (x-a_k)^2 dF_k(x)$$

we find the inequality

$$\mathbf{P} \left\{ \max_{1 \leq k \leq n} |\xi_k - a_k| \geq \tau B_n \right\} \leq \frac{1}{\tau^2 B_n^2} \sum_{k=1}^n \int_{|x-a_k| > \tau B_n} (x-a_k)^2 dF_k(x)$$

By virtue of the Lindeberg condition, the latter sum tends to zero as $n \rightarrow \infty$ for any constant $\tau > 0$. Thus the Lindeberg condition is a peculiar kind of demand for the uniform smallness of the terms $\frac{1}{B_n}(\xi_k - a_k)$ in the sum (2).

Let us note once again that the meaning of the conditions that are sufficient for convergence of the distributions of the sum (2) to the normal law was fully elucidated already in the investigations of A. A. Markov and A. M. Lyapunov.

Sec. 42. Lyapunov's Theorem

We begin by proving the sufficiency of the Lindeberg condition.

Theorem. *If a sequence of mutually independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ for any constant $\tau > 0$ satisfies the Lindeberg condition*

$$\lim_{n \rightarrow \infty} \frac{1}{B_n^2} \sum_{k=1}^n \int_{|x-a_k| > \tau B_n} (x-a_k)^2 dF_k(x) = 0 \quad (1)$$

then as $n \rightarrow \infty$

$$\mathbf{P} \left\{ \frac{1}{B_n} \sum_{k=1}^n (\xi_k - a_k) < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz \quad (2)$$

uniformly in x .

Proof. For brevity we introduce the following notations:

$$\xi_{nk} = \frac{\xi_k - a_k}{B_n},$$

$$F_{nk}(x) = \mathbf{P} \{ \xi_{nk} < x \}$$

It is obvious that

$$\mathbf{M}\xi_{nk} = 0$$

$$\mathbf{D}\xi_{nk} = \frac{1}{B_n^2} \mathbf{D}\xi_k$$

and, consequently,

$$\sum_{k=1}^n \mathbf{D}\xi_{nk} = 1 \quad (2')$$

It is easy to see that in these notations the Lindeberg condition becomes

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n \int_{|x| > \tau} x^2 dF_{nk}(x) = 0 \quad (1')$$

The characteristic function of the sum

$$\frac{1}{B_n} \sum_{k=1}^n (\xi_k - a_k) = \sum_{k=1}^n \xi_{nk}$$

is equal to

$$\varphi_n(t) = \prod_{k=1}^n f_{nk}(t)$$

We have to prove that

$$\lim_{n \rightarrow \infty} \varphi_n(t) = e^{-\frac{t^2}{2}}$$

For this purpose we first establish that the factors $f_{nk}(t)$ tend to 1 uniformly in k ($1 \leq k \leq n$) as $n \rightarrow \infty$. Indeed, taking into account the equation $\mathbf{M}\xi_{nk} = 0$, we find:

$$f_{nk}(t) - 1 = \int (e^{itx} - 1 - itx) dF_{nk}(x)$$

Since for any real α^*

$$|e^{i\alpha} - 1 - i\alpha| \leq \frac{\alpha^2}{2} \quad (3)$$

* This inequality and a whole series of similar ones may be derived, for example, as follows. From the fact that

$$|e^{i\alpha} - 1| = \left| \int_0^\alpha e^{ix} dx \right| \leq \alpha \quad (\alpha > 0)$$

it follows that

$$|f_{nk}(t) - 1| \leq \frac{t^2}{2} \int x^2 dF_{nk}(x)$$

Let ε be an arbitrary positive number; then, clearly,

$$\begin{aligned} \int x^2 dF_{nk}(x) &= \\ &= \int_{|x| \leq \varepsilon} x^2 dF_{nk}(x) + \int_{|x| > \varepsilon} x^2 dF_{nk}(x) \leq \varepsilon^2 + \int_{|x| > \varepsilon} x^2 dF_{nk}(x) \end{aligned}$$

For n sufficiently large, the latter summand may, by (1'), be made less than ε^2 . Thus, for all sufficiently large n and for t in any finite interval $|t| \leq T$,

$$|f_{nk}(t) - 1| \leq \varepsilon^2 T^2$$

uniformly in k ($1 \leq k \leq n$). From this we conclude that

$$\lim_{n \rightarrow \infty} f_{nk}(t) = 1 \quad (4)$$

uniformly in k ($1 \leq k \leq n$) and that for all sufficiently large n , for t lying in an arbitrarily finite interval $|t| \leq T$, the following inequality holds:

$$|f_{nk}(t) - 1| < \frac{1}{2} \quad (5)$$

We can therefore, in the interval $|t| \leq T$ write the expansion (log represents the *principal value* of the logarithm)

$$\begin{aligned} \log \varphi_n(t) &= \sum_{k=1}^n \log f_{nk}(t) = \sum_{k=1}^n \log [1 + (f_{nk}(t) - 1)] = \\ &= \sum_{k=1}^n (f_{nk}(t) - 1) + R_n \quad (6) \end{aligned}$$

we have the inequality

$$|e^{i\alpha} - 1 - i\alpha| = \left| \int_0^\alpha (e^{ix} - 1) dx \right| \leq \frac{\alpha^2}{2}$$

From the latter inequality it follows that

$$\begin{aligned} \left| e^{i\alpha} - 1 - i\alpha + \frac{\alpha^2}{2} \right| &= \left| \int_0^\alpha (e^{ix} - 1 - ix) dx \right| \leq \\ &\leq \int_0^\alpha |e^{ix} - 1 - ix| dx \leq \int_0^\alpha \frac{x^2}{2} dx = \frac{\alpha^3}{6} \quad (3') \end{aligned}$$

and so forth.

where

$$R_n = \sum_{k=1}^n \sum_{s=2}^{\infty} \frac{(-1)^s}{s} (f_{nk}(t) - 1)^s$$

By virtue of (5)

$$\begin{aligned} |R_n| &\leq \sum_{k=1}^n \sum_{s=2}^{\infty} \frac{1}{2} |f_{nk}(t) - 1|^s = \\ &= \frac{1}{2} \sum_{k=1}^n \frac{|f_{nk}(t) - 1|^2}{1 - |f_{nk}(t) - 1|} \leq \sum_{k=1}^n |f_{nk}(t) - 1|^2 \end{aligned}$$

Since

$$\begin{aligned} \sum_{k=1}^n |f_{nk}(t) - 1| &= \sum_{k=1}^n \left| \int (e^{itx} - 1 - itx) dF_{nk}(x) \right| \leq \\ &\leq \frac{t^2}{2} \sum_{k=1}^n \int x^2 dF_{nk}(x) = \frac{t^2}{2} \end{aligned}$$

it follows that

$$|R_n| \leq \frac{t^2}{2} \max_{1 \leq k \leq n} |f_{nk}(t) - 1|$$

From (4) it follows that in an arbitrary finite interval $|t| \leq T$, as n tends to infinity

$$R_n \rightarrow 0 \quad (7)$$

uniformly in t . But

$$\sum_{k=1}^n (f_{nk}(t) - 1) = -\frac{t^2}{2} + \rho_n \quad (8)$$

where

$$\rho_n = \frac{t^2}{2} + \sum_{k=1}^n \int (e^{itx} - 1 - itx) dF_{nk}(x)$$

Let ε be any arbitrary positive number; then by (2')

$$\begin{aligned} \rho_n &= \sum_{k=1}^n \int_{|x| \leq \varepsilon} \left(e^{itx} - 1 - itx - \frac{(itx)^2}{2} \right) dF_{nk}(x) + \\ &\quad + \sum_{k=1}^n \int_{|x| > \varepsilon} \left(\frac{t^2 x^2}{2} + e^{itx} - 1 - itx \right) dF_{nk}(x) \end{aligned}$$

The inequalities (3) and (3') permit obtaining the following estimate:

$$\begin{aligned}
 |\rho_n| &\leq \frac{|t|^3}{6} \sum_{k=1}^n \int_{|x| \leq \varepsilon} |x|^3 dF_{nk}(x) + t^2 \sum_{k=1}^n \int_{|x| > \varepsilon} x^2 dF_{nk}(x) \leq \\
 &\leq \frac{|t|^3}{6} \varepsilon \sum_{k=1}^n \int_{|x| \leq \varepsilon} x^2 dF_{nk}(x) + t^2 \sum_{k=1}^n \int_{|x| > \varepsilon} x^2 dF_{nk}(x) = \\
 &= \frac{|t|^3}{6} \varepsilon + t^2 \left(1 - \frac{|t|}{6} \varepsilon\right) \sum_{k=1}^n \int_{|x| > \varepsilon} x^2 dF_{nk}(x)
 \end{aligned}$$

According to the condition (1'), the second summand may be made less than any $\eta > 0$ for any $\varepsilon > 0$, so long as n is sufficiently large. And since ε is an arbitrary positive number, we can select it so small that, no matter what $\eta > 0$ and T , for all t within the interval $|t| \leq T$, the following inequality will hold:

$$|\rho_n| < 2\eta \quad (n \geq n_0(\varepsilon, \eta, T))$$

This inequality shows that

$$\lim_{n \rightarrow \infty} \rho_n = 0 \quad (9)$$

uniformly in every finite interval of values of t . Collecting together the relations (6), (7), (8) and (9), we finally find that

$$\lim_{n \rightarrow \infty} \log \varphi_n(t) = -\frac{t^2}{2}$$

uniformly in every finite interval of t . The theorem is proved.

Corollary. *If the independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ are identically distributed and have a finite variance different from zero, then as n tends to infinity*

$$\mathbf{P} \left\{ \frac{1}{B_n} \sum_{k=1}^n (\xi_k - \mathbf{M}\xi_k) < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

uniformly in x .

Proof. It suffices to verify that the Lindeberg condition is satisfied under the given assumptions. For this purpose, we note that in our case

$$B_n = b\sqrt{n}$$

where b^2 denotes the variance of a separate summand. Putting $M\xi_k = a$, we can write the following obvious equations:

$$\begin{aligned} \sum_{k=1}^n \frac{1}{B_n^2} \int_{|x-a| > \tau B_n} (x-a)^2 dF_k(x) &= \\ &= \frac{1}{nb^2} n \int_{|x-a| > \tau B_n} (x-a)^2 dF_1(x) = \frac{1}{b^2} \int_{|x-a| > \tau B_n} (x-a)^2 dF_1(x) \end{aligned}$$

From the assumption that the variance is finite and positive we conclude that the integral on the right-hand side of this equation tends to zero as n tends to infinity.

Lyapunov's Theorem. *If for a sequence of mutually independent random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ it is possible to choose a positive number $\delta > 0$ such that as $n \rightarrow \infty$*

$$\frac{1}{B_n^{2+\delta}} \sum_{k=1}^n M|\xi_k - a_k|^{2+\delta} \rightarrow 0 \quad (10)$$

then as n tends to infinity

$$P \left\{ \frac{1}{B_n} \sum_{k=1}^n (\xi_k - a_k) < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

uniformly in x .

Proof. Again, it will suffice to verify that the Lyapunov condition [condition (10)] implies that the Lindeberg condition holds. But this is clear from the following chain of inequalities:

$$\begin{aligned} \frac{1}{B_n^2} \sum_{k=1}^n \int_{|x-a_k| > \tau B_n} (x-a_k)^2 dF_k(x) &\leq \\ &\leq \frac{1}{B_n^2 (\tau B_n)^\delta} \sum_{k=1}^n \int_{|x-a_k| > \tau B_n} |x-a_k|^{2+\delta} dF_k(x) \leq \\ &\leq \frac{1}{\tau^\delta} \frac{\sum_{k=1}^n \int |x-a_k|^{2+\delta} dF_k(x)}{B_n^{2+\delta}} \end{aligned}$$

Sec. 43. The Local Limit Theorem

We shall now indicate the sufficient conditions for application of the other classical limit theorem: the *local theorem*. In doing so, we will confine ourselves to considering only the case of mutually independent summands having one and the same probability distribution.

Let us agree to say that a discrete random variable ξ has a *lattice distribution* if there exist numbers a and $h > 0$ such that all possible values of ξ may be represented in the form $a + kh$, where the parameter k can assume any integral values ($-\infty < k < \infty$).

The Poisson, Bernoulli and other distributions are lattice distributions.

Let us now express the lattice nature of a distribution of a random variable ξ in terms of characteristic functions. For this purpose we prove the following lemma.

Lemma. *For a random variable ξ to have a lattice distribution it is necessary and sufficient that for some $t \neq 0$ the absolute value of its characteristic function be equal to unity.*

Proof. Indeed, if ξ is lattice-distributed and p_k is the probability of the equation $\xi = a + kh$, then the characteristic function of the variable ξ is equal to

$$f(t) = \sum_{k=-\infty}^{\infty} p_k e^{it(a+kh)} = e^{iat} \sum_{k=-\infty}^{\infty} p_k e^{itkh}$$

From here we find

$$f\left(\frac{2\pi}{h}\right) = e^{2\pi i \frac{a}{h}} \sum_{k=-\infty}^{\infty} p_k e^{2\pi i k} = e^{2\pi i \frac{a}{h}}$$

We thus see that for every lattice distribution

$$\left| f\left(\frac{2\pi}{h}\right) \right| = 1$$

Now suppose that for some $t_1 \neq 0$

$$|f(t_1)| = 1$$

and prove that the variable ξ then has a lattice distribution. The last equation implies that for some θ

$$f(t_1) = e^{i\theta}$$

Thus,

$$\int e^{it_1 x} dF(x) = e^{i\theta}$$

and, consequently,

$$\int e^{i(t_1 x - \theta)} dF(x) = 1$$

From this it follows that

$$\int \cos(t_1 x - \theta) dF(x) = 1$$

For this equation to be possible, it is necessary that the function $F(x)$ be allowed to grow only for those values of x for which

$$\cos(t_1 x - \theta) = 1$$

This means that the possible values of ξ must be of the form

$$x = \frac{\theta}{t_1} + k \frac{2\pi}{t_1}$$

Q.E.D.

We will call the number h the *span of the distribution*. The distribution span h is a *maximum* if, no matter what the choice of b ($-\infty < b < \infty$) and $h_1 > h$, it is impossible to represent all possible values of ξ in the form $b + kh_1$.

To illustrate the difference between the concepts of distribution span and maximal distribution span, we consider the following example. Let ξ assume all odd numbers as its values. Obviously, all values of ξ may be written as $a + kh$, where $a = 0$ and $h = 1$. The span h cannot, however, be maximal, since all possible values of ξ may also be written as $b + kh_1$, where $b = 1$ and $h_1 = 2$.

The conditions for a distribution span to be maximal may be expressed otherwise.

Firstly, the span h may be maximal if and only if the greatest common divisor of paired differences of possible values of the variable ξ divided by h is equal to unity.

Secondly, the span h is maximal if and only if the absolute value of the characteristic function is less than unity in the interval $0 < |t| < \frac{2\pi}{h}$ and is equal to unity when $t = \frac{2\pi}{h}$.

The latter assertion is a straightforward consequence of the lemma that has just been proved. Indeed, if for $0 < t_1 < \frac{2\pi}{h}$

$$|f(t_1)| = 1$$

then according to the proof the quantity $\frac{2\pi}{t_1}$ must be the distribution span, and since

$$h < \frac{2\pi}{t_1}$$

the span h cannot be maximal.

From this we can conclude that if h is a maximal distribution span, then for each $\varepsilon > 0$ there will be a number $c_0 > 0$ such that for all t in the interval $\varepsilon \leq |t| \leq \frac{2\pi}{h} - \varepsilon$ the following inequality will hold:

$$|f(t)| \leq e^{-c_0} \quad (1)$$

Now let the random variables $\xi_1, \xi_2, \dots, \xi_n, \dots$ be mutually independent, lattice-distributed and have one and the same distribution function $F(x)$. Consider the sum

$$\zeta_n = \xi_1 + \xi_2 + \dots + \xi_n$$

It is obviously also a lattice random variable and its possible values can be written as $na + kh$. Denote by $P_n(k)$ the probability of the equation

$$\zeta_n = na + kh$$

in particular, $P_1(k) = \mathbf{P}\{\xi_1 = a + kh\} = p_k$.

Further denote

$$z_{nk} = \frac{an + kh - A_n}{B_n}$$

where $A_n = \mathbf{M}\zeta_n$, $B_n^2 = \mathbf{D}\zeta_n = n\mathbf{D}\xi_1$.

We can now prove the following proposition which in obvious fashion generalizes the local limit theorem of DeMoivre-Laplace.

Theorem*. *Let the independent lattice random variables*

$$\xi_1, \xi_2, \dots, \xi_n, \dots$$

have one and the same distribution function $F(x)$ and let their expectations and variances be finite. Then for the relation

$$\frac{B_n}{h} P_n(k) - \frac{1}{\sqrt{2\pi}} e^{-\frac{z_{nk}^2}{2}} \rightarrow 0$$

to hold and be uniform in k ($-\infty < k < \infty$) as n tends to infinity, it is necessary and sufficient that the distribution span h be maximal.

Proof. The necessity of the hypothesis is almost obvious. Indeed, if the span h is not maximal, then the possible values of the sum

$\zeta_n = \sum_{k=1}^n \xi_k$ will have systematic omissions: the difference between the closest possible values of the sum cannot be less than dh , where d is the greatest common divisor of the differences of possible values of ζ_n divided by h . If h is not the maximal span, then $d > 1$ for all values of n .

Proof of the sufficiency of the hypothesis requires a somewhat more complicated argument.

The characteristic function of the variable ξ_k ($k = 1, 2, 3, \dots$) is

$$f(t) = \sum_{k=-\infty}^{\infty} p_k e^{iat + itkh} = e^{iat} \sum_{k=-\infty}^{\infty} p_k e^{itkh}$$

* This theorem was proved by B. V. Gnedenko.—Ed.

and the characteristic function of the sum ζ_n is

$$f^n(t) = e^{ian t} \sum_{k=-\infty}^{\infty} P_n(k) e^{itkh}$$

Multiplying the last equation by $e^{-ian t - itkh}$ and integrating it from $-\frac{\pi}{h}$ to $\frac{\pi}{h}$, we get

$$\frac{2\pi}{h} P_n(k) = \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} f^n(t) e^{-ian t - itkh} dt$$

Noting that

$$hk = B_n z_{nk} + A_n - an$$

(we will write z in place of z_{nk}), we can write

$$\frac{2\pi}{h} P_n(k) = \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} f^{*n}(t) e^{-itz B_n} dt$$

where

$$f^*(t) = e^{-\frac{it A_n}{n}} f(t)$$

Finally, putting $x = t B_n$, we obtain

$$\frac{2\pi B_n}{h} P_n(k) = \int_{-\frac{\pi B_n}{h}}^{\frac{\pi B_n}{h}} e^{-izx} f^{*n}\left(\frac{x}{B_n}\right) dx$$

It is easy to calculate that

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = \frac{1}{2\pi} \int e^{-izx - \frac{x^2}{2}} dx$$

Let us represent the difference

$$R_n = 2\pi \left[\frac{B_n}{h} P_n(k) - \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \right]$$

in the form of a sum of four integrals:

$$R_n = J_1 + J_2 + J_3 + J_4$$

where

$$\begin{aligned} J_1 &= \int_{-A}^A e^{-izx} \left[f^{*n} \left(\frac{x}{B_n} \right) - e^{-\frac{x^2}{2}} \right] dx \\ J_2 &= - \int_{|x| > A} e^{-izx - \frac{x^2}{2}} dx \\ J_3 &= \int_{\varepsilon B_n \leq |x| \leq \frac{\pi B_n}{h}} e^{-izx} f^{*n} \left(\frac{x}{B_n} \right) dx \\ J_4 &= \int_{A \leq |x| < \varepsilon B_n} e^{-izx} f^{*n} \left(\frac{x}{B_n} \right) dx \end{aligned}$$

where $A > 0$ is a sufficiently large constant and $\varepsilon > 0$ is a sufficiently small constant, the more precise values of which will be chosen later on.

By virtue of the corollary to the theorem proved in the preceding section, in any finite interval of values of t the relation

$$f^{*n} \left(\frac{t}{B_n} \right) \rightarrow e^{-\frac{t^2}{2}} \quad (n \rightarrow \infty)$$

holds uniformly in t . But from this it follows that whatever the constant A ,

$$J_1 \rightarrow 0 \quad (n \rightarrow \infty)$$

The integral J_2 is estimated by means of the inequality

$$|J_2| \leq \int_{|x| > A} e^{-\frac{x^2}{2}} dx \leq \frac{2}{A} \int_A^\infty x e^{-\frac{x^2}{2}} dx = \frac{2}{A} e^{-\frac{A^2}{2}}$$

Choosing A sufficiently large, we can make J_2 as small as we like.

By the inequality (1) we have

$$|J_3| \leq \int_{\varepsilon B_n \leq |x| \leq \frac{\pi B_n}{h}} \left| f^* \left(\frac{x}{B_n} \right) \right|^n dx \leq e^{-nc_0} 2B_n \left(\frac{\pi}{h} - \varepsilon \right)$$

Whence it is clear that as n tends to infinity

$$J_3 \rightarrow 0$$

To estimate the integral J_4 , we note that the existence of variance implies the existence of the second derivative of the function $f^*(t)$. We can therefore, in accordance with (3) of Sec. 35, make use of the expansion

$$f^*(t) = 1 - \frac{\sigma^2 t^2}{2} + o(t^2)$$

in the neighbourhood of the point $t=0$; and for $|t| \leq \varepsilon$, if ε is sufficiently small, we get

$$|f^*(t)| < 1 - \frac{\sigma^2 t^2}{4} < e^{-\frac{\sigma^2 t^2}{4}}$$

Then, for $|x| \leq \varepsilon B_n$,

$$\left| f^* \left(\frac{x}{B_n} \right) \right|^n > e^{-\frac{n\sigma^2 t^2}{4B_n^2}} = e^{-\frac{t^2}{4}}$$

And so

$$|J_4| \leq 2 \int_A^{\varepsilon B_n} e^{-\frac{t^2}{4}} dt < 2 \int_A^{\infty} e^{-\frac{t^2}{4}} dt$$

By choosing A sufficiently large we can make the integral J_4 as small as we desire. The theorem is proved.

There is yet another case when it is natural to pose the question of the local behaviour of the distribution functions of sums. This is the case of continuous distributions.

The question is: When do the density functions of normalized sums converge to the normal density function if the corresponding distribution functions converge to the normal distribution? This problem is exhaustively solved in the following theorem.

Theorem. *Let the independent random variables*

$$\xi_1, \xi_2, \dots, \xi_n, \dots$$

have one and the same distribution function $F(x)$; their expectations and variances are finite and, beginning with a certain n_0 , the random variable

$$s_n = \frac{1}{\sqrt{nD\xi_1}} \sum_{k=1}^n (\xi_k - M\xi_k)$$

has a density function $p_n(x)$. So that as n tends to infinity

$$p_n(x) - \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \rightarrow 0$$

uniformly in x ($-\infty < x < \infty$), it is necessary and sufficient that there exist a number n_1 such that the function $p_{n_1}(x)$ is bounded.

We shall not give the proof of this theorem since it repeats to a great extent the reasoning that has just been given and rests on Theorem 3 of Sec. 36 and the lemma of Sec. 39.

EXERCISES

1. Prove that as n tends to infinity

$$\frac{1}{\Gamma\left(\frac{n}{2}\right)} \sqrt{\left(\frac{n}{2}\right)^n} \int_0^{\sqrt{\frac{2}{n}}} z^{\frac{n}{2}-1} e^{-\frac{nz}{2}} dz \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{1}{2}z^2} dz$$

Hint. Apply Lyapunov's theorem to the chi-square distribution.

2. The random variables

$$\xi_n = \begin{cases} -n^\alpha & \text{with probability } \frac{1}{2} \\ +n^\alpha & \text{with probability } \frac{1}{2} \end{cases}$$

are independent. Prove that for $\alpha > -\frac{1}{2}$ the Lyapunov theorem may be applied to them

3. Prove that as $n \rightarrow \infty$,

$$e^{-n} \sum_{k=0}^n \frac{n^k}{k!} \rightarrow \frac{1}{2}$$

Hint. Apply the Lyapunov theorem to the sum of the Poisson distributed random variables with parameter $\lambda=1$.

4. The probability of occurrence of an event A in the i th trial is equal to p_i ; μ is the number of occurrences of A in n independent trials. Prove that

$$P \left\{ \frac{\mu - \sum_{k=1}^n p_k}{\sqrt{\sum_{i=1}^n p_i q_i}} < x \right\} \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{z^2}{2}} dz$$

if and only if $\sum_{i=1}^{\infty} p_i q_i = \infty$.

5. Prove that under the conditions of the preceding problem, the requirement that $\sum_{i=1}^{\infty} p_i q_i = +\infty$ is sufficient not only for the integral theorem but for the local theorem as well.

The Theory of Infinitely Divisible Distribution Laws

For a long time the central problem of probability theory was considered to be the finding of the most general conditions under which the distribution functions of sums of independent random variables converge to the normal law. Extremely general conditions sufficient for this convergence were found by A. M. Lyapunov (see Chap. 8). Attempts to expand Lyapunov's conditions were successful only in recent years when conditions were found that are not only sufficient but also—under extremely natural restrictions—necessary.

In parallel with the consummation of the classical range of problems, there developed a new trend in the theory of limit theorems for the sums of independent random variables closely associated with the introduction and development of the theory of stochastic (random) processes. The first question to arise was: What laws, in addition to the normal law, may be limit laws for sums of independent random variables?

It was found that the class of limit laws is not exhausted by far by the normal law. Then the question arose of defining the conditions that must be imposed on the summands so that the distribution functions of the sums converged to one or another limit law.

In the present chapter our purpose is to describe some investigations of recent years devoted to limit theorems for sums of independent random variables. Here, we confine ourselves to the case when the summands have finite variances. Consideration of the problem without this restriction demands more cumbersome calculations; we refer the interested reader to its solution in the monograph by Gnedenko and Kolmogorov that was mentioned earlier. As a simple consequence of the general theorems that we present, we will obtain the earlier mentioned necessary and sufficient condition for the convergence of the distribution functions of sums to the normal law.

Sec. 44. Infinitely Divisible Laws and Their Basic Properties

A distribution law $\Phi(x)$ is called *infinitely divisible* if, no matter what natural number n is taken, the random variable distributed in accordance with the $\Phi(x)$ law is the sum of n independent random variables $\xi_1, \xi_2, \dots, \xi_n$ with one and the same distribution law $\Phi_n(x)$ (dependent on number of summands n).

It is clear that this definition is equivalent to the following: the law $\Phi(x)$ is called infinitely divisible if for any n its characteristic function is the n th power of some other characteristic function.

Researches in recent years have shown that infinitely divisible laws play a significant role in a variety of problems of probability theory. For one thing, it has turned out that the class of limit laws for sums of independent random variables coincides with the class of infinitely divisible laws.

We now take up the properties of infinitely divisible laws that will be needed later on. We begin with the proof that the normal and the Poisson laws are infinitely divisible. Indeed, the characteristic function of the normal law with expectation a and variance σ^2 is equal to

$$\varphi(t) = e^{iat - \frac{1}{2} \sigma^2 t^2}$$

For any n , the n th root of $\varphi(t)$ is again the characteristic function of a normal law, but with expectation $\frac{a}{n}$ and variance $\frac{\sigma^2}{n}$.

We will generalize somewhat the earlier encountered concept of Poisson's law and we will say that a random variable ξ is Poisson distributed if it can assume only values $ak + b$, where a and b are real constants, and $k = 0, 1, 2, \dots$, and

$$\mathbf{P} \{ \xi = ak + b \} = \frac{e^{-\lambda} \lambda^k}{k!} \quad (1)$$

where λ is a positive constant. It is easy to calculate that the characteristic function for the law (1) is given by the formula

$$\varphi(t) = e^{\lambda(e^{iat} - 1) + ibt}$$

We see that for any n , the n th root of $\varphi(t)$ is again the characteristic function of Poisson's law but with different parameters: a , $\frac{\lambda}{n}$ and $\frac{1}{n}b$.

Theorem 1. *The characteristic function of an infinitely divisible law does not vanish.*

Proof. Let $\Phi(x)$ be an infinitely divisible law and $\varphi(t)$ its characteristic function. Then, by definition, for any n we have the equation

$$\varphi(t) = \{\varphi_n(t)\}^n \quad (2)$$

where $\varphi_n(t)$ is some characteristic function. By virtue of the continuity of the function $\varphi(t)$ there exists a range of values of the argument $|t| \leq a$, in which $\varphi(t) \neq 0$; clearly, in this same region $\varphi_n(t) \neq 0$. For n sufficiently large, we can make the quantity $|\varphi_n(t)| = \sqrt[n]{|\varphi(t)|}$ arbitrarily close to unity uniformly in t ($|t| \leq a$).

Now take two mutually independent random variables η_1 and η_2 distributed in accordance with some law $F(x)$ and consider their difference $\eta = \eta_1 - \eta_2$. The characteristic function of the variable η is

$$f^*(t) = \mathbf{M}e^{it(\eta_1 - \eta_2)} = |\mathbf{M}e^{it\eta_1}|^2 = |f(t)|^2$$

We thus see that the square of the absolute value of any characteristic function is a characteristic function.

Further, since a real characteristic function is of the form

$$f(t) = \int \cos xt \, dF(x)$$

it therefore follows that we can write the inequality

$$\begin{aligned} 1 - f(2t) &= \int (1 - \cos 2xt) \, dF(x) = \\ &= 2 \int \sin^2 xt \, dF(x) = 2 \int (1 - \cos xt)(1 + \cos xt) \, dF(x) \leq \\ &\leq 4 \int (1 - \cos xt) \, dF(x) = 4(1 - f(t)) \end{aligned}$$

From the foregoing we see that the function $|\varphi_n(t)|^2$ satisfies the inequality

$$1 - |\varphi_n(2t)|^2 \leq 4(1 - |\varphi_n(t)|^2)$$

From this inequality it follows that if n is so large that $1 - |\varphi_n(t)| < \varepsilon$ for $|t| \leq a$, then in this region

$$1 - |\varphi_n(2t)| \leq 1 - |\varphi_n(2t)|^2 \leq 4(1 - |\varphi_n(t)|^2) \leq 8(1 - |\varphi_n(t)|) < 8\varepsilon$$

Summarizing, in the region $|t| \leq 2a$

$$1 - |\varphi_n(t)| < 8\varepsilon$$

Thus, for n sufficiently large in the region of $|t| \leq 2a$, $\varphi_n(t)$ and also $\varphi(t)$ do not vanish.

In similar fashion we prove that $\varphi(t) \neq 0$ in the region $|t| < 4a$, and so on.

This proves our theorem.

Theorem 2. *The distribution function of a sum of independent random variables having infinitely divisible distribution functions is also infinitely divisible.*

Proof. It is obviously sufficient to confine oneself to the case of two summands in order to prove the theorem. If $\varphi(t)$ and $\psi(t)$ are the characteristic functions of the summands, then by hypothesis we have, for any n ,

$$\varphi(t) = \{\varphi_n(t)\}^n, \psi(t) = \{\psi_n(t)\}^n$$

where $\varphi_n(t)$ and $\psi_n(t)$ are characteristic functions. Therefore, the characteristic function of a sum, for any n , satisfies the equation

$$\chi(t) = \varphi(t) \cdot \psi(t) = \{\varphi_n(t) \cdot \psi_n(t)\}^n$$

Theorem 3. *The limit distribution function (in the meaning of being weakly convergent) of a sequence of infinitely divisible distribution functions is itself infinitely divisible.*

Proof. Let the sequence $\Phi^{(k)}(x)$ of infinitely divisible distribution functions be weakly convergent to the distribution function $\Phi(x)$. Then

$$\lim_{k \rightarrow \infty} \varphi^{(k)}(t) = \varphi(t) \quad (3)$$

uniformly in each finite interval t . By hypothesis, for any n , the functions $\sqrt[n]{\varphi^{(k)}(t)}$ (is understood to be its principal value)

$$\varphi_n^{(k)}(t) = \sqrt[n]{\varphi^{(k)}(t)} \quad (4)$$

are characteristic functions. From (3) we conclude that for every n

$$\lim_{k \rightarrow \infty} \varphi_n^{(k)}(t) = \varphi_n(t) \quad (5)$$

The continuity of $\varphi_n(t)$ follows from the continuity of $\varphi_n^{(k)}(t)$. By virtue of the limit theorem for characteristic functions, $\varphi_n(t)$ is a characteristic function. From (3), (4) and (5) we find that for every n we have the equation

$$\varphi(t) = \{\varphi_n(t)\}^n$$

Q.E.D.

Sec. 45. The Canonical Representation of Infinitely Divisible Laws

From now on we confine ourselves to the study of infinitely divisible laws with *finite variance*. The purpose of this section is to prove the following theorem, which was found in 1932 by A. N. Kolmogorov and which gives a complete description of the class of distribution laws that interests us.

Theorem. *For a distribution function $\Phi(x)$ with finite variance to be infinitely divisible, it is necessary and sufficient that the logarithm of*

its characteristic function have the form

$$\log \varphi(t) = i\gamma t + \int \{e^{itx} - 1 - itx\} \frac{1}{x^2} dG(x) \quad (1)$$

where γ is a real constant and $G(x)$ is a nondecreasing function of bounded variation.

Proof. First assume that $\Phi(x)$ is an infinitely divisible law and $\varphi(t)$ is its characteristic function. Then for any n

$$\varphi(t) = \{\varphi_n(t)\}^n$$

where $\varphi_n(t)$ is some characteristic function. Since $\varphi(t) \neq 0$, this equation is equivalent to the following one:

$$\log \varphi(t) = n \log \varphi_n(t) = n \log[1 + (\varphi_n(t) - 1)]$$

For any T , as n tends to infinity,

$$\varphi_n(t) \rightarrow 1$$

uniformly in the interval $|t| < T$; for this reason, in any finite interval of values of t the quantity $|\varphi_n(t) - 1|$ may be made less than any preassigned number so long as n is sufficiently great. We can therefore take advantage of the equation

$$\log[1 + (\varphi_n(t) - 1)] = (\varphi_n(t) - 1)(1 + o(1))$$

which yields

$$\log \varphi(t) = \lim_{n \rightarrow \infty} n(\varphi_n(t) - 1) = \lim_{n \rightarrow \infty} n \int (e^{itx} - 1) d\Phi_n(x) \quad (2)$$

where $\Phi_n(x)$ is a distribution function having $\varphi_n(t)$ as its characteristic function. From the definition of expectation and from the relation between the functions $\Phi_n(x)$ and $\Phi(x)$ it follows that

$$n \int x d\Phi_n(x) = \int x d\Phi(x)$$

We denote this quantity by γ ; then Equation (2) may be rewritten in the following form:

$$\log \varphi(t) = i\gamma t + \lim_{n \rightarrow \infty} n \int \{e^{itx} - 1 - itx\} d\Phi_n(x)$$

Now put

$$G_n(x) = n \int_{-\infty}^x u^2 d\Phi_n(u)$$

Obviously, the functions $G_n(x)$ do not decrease with increasing argument and $G_n(-\infty) = 0$. Besides, the functions $G_n(x)$ are uniformly bounded. The last assertion follows from the properties of variance and the relations between the functions $\Phi(x)$ and $\Phi_n(x)$.

Indeed,

$$G_n(+\infty) = n \int u^2 d\Phi_n(u) = n \left[\int u^2 d\Phi_n(u) - \left(\int u d\Phi_n(u) \right)^2 \right] + \\ + n \left(\int u d\Phi_n(u) \right)^2 = \sigma^2 + \frac{1}{n} \gamma^2 \quad (3)$$

where σ^2 is the variance of the law $\Phi(x)$.

In the new notations (see Property 6 of the Stieltjes integral in Sec. 25),

$$\log \varphi(t) = i\gamma t + \lim_{n \rightarrow \infty} \int (e^{itx} - 1 - itx) \frac{1}{x^2} dG_n(x)$$

By Helly's first theorem, from the sequence of functions $G_n(x)$ one can choose a subsequence that converges to some limit function $G(x)$. If $A < 0$ and $B > 0$ are continuity points of the functions $G(x)$, then by virtue of the second theorem of Helly, as $k \rightarrow \infty$,

$$\int_A^B (e^{itx} - 1 - itx) \frac{1}{x^2} dG_{n_k}(x) \rightarrow \int_A^B (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x) \quad (4)$$

We know that

$$|e^{itx} - 1 - itx| \leq |e^{itx} - 1| + |tx| \leq |tx| + |tx| = 2|t| \cdot |x|$$

and so

$$\left| \int_{-\infty}^A + \int_B^{\infty} (e^{itx} - 1 - itx) \frac{1}{x^2} dG_{n_k}(x) \right| \leq \int_{-\infty}^A + \int_B^{\infty} \frac{|e^{itx} - 1 - itx|}{x^2} dG_{n_k}(x) \leq \\ \leq 2|t| \left(\int_{-\infty}^A + \int_B^{\infty} \frac{1}{|x|} dG_{n_k}(x) \right) \leq \frac{2|t|}{\Gamma} \left(\int_{-\infty}^A + \int_B^{\infty} dG_{n_k}(x) \right) \leq \\ \leq \frac{2|t|}{\Gamma} \max_{1 \leq k < \infty} \int dG_{n_k}(x)$$

where $\Gamma = \min(|A|, B)$. Since the variations of the functions $G_{n_k}(u)$ are uniformly bounded, for any $\varepsilon > 0$, we can—by choosing A and B sufficiently large—make the following inequality hold:

$$\left| \int_{-\infty}^A + \int_B^{\infty} (e^{itx} - 1 - itx) \frac{1}{x^2} dG_{n_k}(x) \right| < \frac{\varepsilon}{2} \quad (5)$$

for all t contained in some finite interval, and for all k .

From (4) and (5) it follows that for any $\varepsilon > 0$, for all t contained in an arbitrary finite interval, given sufficiently large n , the following inequality holds:

$$\left| \int (e^{itx} - 1 - itx) \frac{1}{x^2} dG_{n_k}(x) - \int (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x) \right| < \varepsilon.$$

in other words,

$$\lim_{k \rightarrow \infty} \int (e^{itx} - 1 - itx) \frac{1}{x^2} dG_{n_k}(x) = \int (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x)$$

We have thus proved that the logarithm of the characteristic function of any infinitely divisible law may be written in the form of (1). Now we have to prove the converse proposition, that any function whose logarithm is expressible by formula (1) is the characteristic function of some infinitely divisible law.

For any ε ($0 < \varepsilon < 1$) the integral

$$\int_{\varepsilon}^{\frac{1}{\varepsilon}} (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x) \quad (6)$$

by definition of the Stieltjes integral, is the limit of the sums

$$\sum_{s=1}^n (e^{it\bar{x}_s} - 1 - it\bar{x}_s) \frac{1}{x_s^2} (G(x_{s+1}) - G(x_s))$$

where $x_1 = \varepsilon$, $x_{n+1} = \frac{1}{\varepsilon}$, $x_s \leq \bar{x}_s \leq x_{s+1}$ and $\max(x_{s+1} - x_s) \rightarrow 0$.

Each term of this sum is the logarithm of the characteristic function of some Poisson law. According to Theorems 2 and 3 of Sec. 44, the integral (6) is the logarithm of the characteristic function of some infinitely divisible law. Passing to the limit as $\varepsilon \rightarrow 0$, we convince ourselves that we have the very same thing for the integral

$$\int_{x>0} (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x) \quad (7)$$

In similar fashion we prove that the integral

$$\int_{x<0} (e^{itx} - 1 - itx) \frac{1}{x^2} dG(x) \quad (8)$$

is the logarithm of the characteristic function of some infinitely divisible law. The integral on the right-hand side of formula (1) is equal to the sum of the integrals (7) and (8) and the quantity

$$i\gamma t - \frac{1}{2} t^2 (G(+0) - G(-0))$$

This last term is the logarithm of the characteristic function of the normal law. From Theorem 2, Sec. 44, it follows that the function $\phi(t)$, expressible by means of (1), is the characteristic

function of some infinitely divisible law.* It now remains for us to convince ourselves that the representation of $\log \varphi(t)$ by (1) is unique, i.e., that the function $G(x)$ and the constant γ are uniquely determined by *specification* of $\varphi(t)$.

By differentiating formula (1) we find

$$\frac{d^2}{dt^2} \log \varphi(t) = - \int e^{itx} dG(x) \quad (9)$$

From the theory of characteristic functions we know that the function $G(x)$ in this formula is uniquely determined by $\frac{d^2}{dt^2} \log \varphi(t)$. While proving the theorem we saw that the constant γ is the expectation and, hence, is also uniquely determined by the function $\varphi(t)$.

Finally, we note the probabilistic meaning of the total variation of the function $G(x)$. We know that if a random variable ξ is distributed according to the law $\Phi(x)$, then (see (5) of Sec. 35)

$$D\xi = - \left[\frac{d^2}{dt^2} \log \varphi(t) \right]_{t=0}$$

From (9) it therefore follows that

$$D\xi = \int dG(x) = G(+\infty)$$

The canonical representation of the normal law and the Poisson law may serve as an illustration.

For the normal law with variance σ^2 and expectation a ,

$$\gamma = a \text{ and } G(x) = \begin{cases} 0 & \text{for } x < 0 \\ \sigma^2 & \text{for } x > 0 \end{cases}$$

Indeed, this function and the constant γ lead to this law, since

$$\int \{e^{itx} - 1 - itx\} \frac{1}{x^2} dG(x) = \lim_{u \rightarrow 0} \frac{e^{itu} - 1 - itu}{u^2} [G(+0) - G(-0)] = - \frac{t^2 \sigma^2}{2}$$

and by virtue of the uniqueness of the canonical representation, the other functions $G(x)$ cannot yield the normal law.

In similar manner it is easy to see that to the Poisson law with characteristic function

$$\varphi(t) = e^{\lambda (e^{ita} - 1) + ibt}$$

* We have just proved that any infinitely divisible law is either a convolution of a finite number of Poisson laws and the normal law or the limit of a uniformly converging sequence of such laws. We thus see that the normal and Poisson laws are the basic elements that comprise every infinitely divisible law.

there corresponds the function $G(x)$ with a single jump at the point a :

$$G(x) = \begin{cases} 0 & \text{for } x < a \\ a^2\lambda & \text{for } x > a \end{cases}$$

and $\gamma = b + a\lambda$.

Sec. 46. A Limit Theorem for Infinitely Divisible Laws

We know that if a sequence of infinitely divisible distribution laws converges to a limit distribution law, then this limit law is itself infinitely divisible. We now point out the conditions that suffice for a given sequence of infinitely divisible distribution functions to converge to the limit distribution function.

Theorem. *In order for a sequence $\{\Phi_n(x)\}$ of infinitely divisible distribution functions to converge, as $n \rightarrow \infty$, to some distribution function $\Phi(x)$ and for their variances to converge to the variance of the limit law, it is necessary and sufficient that there exist a constant γ and the function $G(x)$, for which, as $n \rightarrow \infty$,*

- (1) $G_n(x)$ converges weakly to $G(x)$,
- (2) $G_n(\infty) - G_n(-\infty) \rightarrow G(\infty) - G(-\infty)$,
- (3) $\gamma_n \rightarrow \gamma$,

where γ_n and $G_n(x)$ are defined by formula (1), Sec. 45, for the law $\Phi_n(x)$, and the constant γ and the function $G(x)$ define, by the same formula, the limit law $\Phi(x)$.

Proof. The sufficiency of the conditions of the theorem is a direct consequence of Helly's second theorem. Indeed, from the conditions of the theorem and from formula (1), Sec. 45, it follows that as $n \rightarrow \infty$

$$\log \varphi_n(t) \rightarrow \log \varphi(t)$$

uniformly in every finite interval t .

In the preceding section we saw that the integrals

$$\int dG_n(u) \text{ and } \int dG(u)$$

were equal to the variances of the laws $\Phi_n(x)$ and $\Phi(x)$; therefore, the second condition of the theorem is nothing other than the requirement of convergence of the variances.

Suppose we now know that as n tends to infinity

$$\Phi_n(x) \rightarrow \Phi(x) \tag{1}$$

and the variances of the laws $\Phi_n(x)$ converge to the variance of the limit law $\Phi(x)$. We shall prove that these requirements imply fulfillment of the conditions of the theorem. As we have already noticed,

this does not require any supplementary argument with regard to condition 2. From this it follows that the total variations of the functions $G_n(u)$ are uniformly bounded. We can therefore take advantage of Helly's first theorem and from the sequence of functions $G_n(u)$ we can choose a subsequence $G_{n_k}(u)$ that converges to some limit function $G_\infty(u)$ as $k \rightarrow \infty$. Our purpose consists in proving the equation

$$G_\infty(u) = G(u)$$

To do this, we first establish that

$$\begin{aligned} J_k &= \int \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_{n_k}(u) \rightarrow \\ &\rightarrow J_\infty = \int \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_\infty(u) \end{aligned} \quad (2)$$

as k tends to infinity. Let $A < 0$ and $B > 0$ be continuity points of the functions $G_\infty(u)$; then by Helly's second theorem, as $k \rightarrow \infty$,

$$\int_A^B \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_{n_k}(u) \rightarrow \int_A^B \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_\infty(u) \quad (3)$$

On the other hand, from the inequality

$$|e^{itx} - 1 - itx| \leq 2|tx|$$

we see that

$$\begin{aligned} L_k &= \left| \int_{-\infty}^A + \int_B^\infty \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_{n_k}(u) \right| \leq \\ &\leq 2|t| \left| \int_{-\infty}^A + \int_B^\infty \frac{1}{|u|} dG_{n_k}(u) \right| \leq \frac{2|t|}{\Gamma} \left(\int_{-\infty}^A + \int_B^\infty dG_{n_k}(u) \right) \leq \frac{2|t|}{\Gamma} \int dG_{n_k}(u) \end{aligned}$$

where $\Gamma = \min(-A, B)$. By virtue of the uniform boundedness of the variations of the functions $G_n(u)$, for any $\varepsilon > 0$ it is possible to select A and B so large in absolute value that

$$L_k < \varepsilon \quad (4)$$

Similarly, for any $\varepsilon > 0$, the inequality

$$\left| \int_{-\infty}^A + \int_B^\infty \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_\infty(u) \right| < \varepsilon \quad (5)$$

holds for A and B sufficiently large in absolute value. From the relations (3), (4) and (5) we conclude that no matter what $\varepsilon > 0$, for sufficiently large values of k

$$|J_k - J_\infty| < 3\varepsilon$$

Relation (2) is thus proved. From (1) we see that

$$\begin{aligned} \lim_{n \rightarrow \infty} \log \varphi_n(t) &= \lim_{n \rightarrow \infty} \left(i\gamma_n t + \int \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG_n(u) \right) = \\ &= \log \varphi(t) = i\gamma t + \int \{e^{itu} - 1 - itu\} \frac{1}{u^2} dG(u), \end{aligned}$$

or

$$\begin{aligned} \lim_{n \rightarrow \infty} \left(i\gamma_{n_k} + \int \{e^{itu} - 1 - itu\} \frac{1}{tu^2} dG_{n_k}(u) \right) &= \\ &= i\gamma + \int \{e^{itu} - 1 - itu\} \frac{1}{tu^2} dG(u) \end{aligned} \quad (6)$$

From the inequality

$$|e^{itu} - 1 - itu| \leq \frac{t^2 u^2}{2}$$

and the uniform boundedness of the total variations of the functions $G_{n_k}(u)$ we conclude that as $t \rightarrow 0$

$$\left| \int \{e^{itu} - 1 - itu\} \frac{1}{tu^2} dG_{n_k}(u) \right| \leq \left| t \int dG_{n_k}(u) \right| \rightarrow 0$$

uniformly in n . And so, as $t \rightarrow 0$, formula (6) yields

$$\lim_{k \rightarrow \infty} \gamma_{n_k} = \gamma \quad (7)$$

and, on the other hand, from (2) and (7),

$$\log \varphi(t) = i\gamma t + \int \{e^{iut} - 1 - iut\} \frac{1}{u^2} dG_\infty(u)$$

By virtue of the uniqueness of the representation of infinitely divisible laws by formula (1), Sec. 45, we conclude that $G_\infty(u) = G(u)$.

To summarize, then, any convergent sequence of functions $G_{n_k}(u)$ converges to the function $G(u)$, and at the same time the constants γ_{n_k} converge to γ .

It is now easy to prove that the entire sequence $G_n(u)$ also converges to $G(u)$ and, hence, at the same time $\lim_{n \rightarrow \infty} \gamma_n = \gamma$, for otherwise there would be a point of continuity of the functions $G(u)$, call it c , and a subsequence of the functions $G_{n_k}(u)$, which at the point $u=c$ converges to a number different from $G(c)$ as $k \rightarrow \infty$. By Helly's first theorem we can extract from this sequence a convergent subsequence $G_{n_{k_r}}(u)$.

From the foregoing it follows that at all points of continuity of the function $G(u)$

$$\lim_{r \rightarrow \infty} G_{n_{k_r}}(u) = G(u)$$

This contradicts our assumption. Thus, at all points of continuity of the function $G(u)$

$$\lim_{n \rightarrow \infty} G_n(u) = G(u)$$

From this, as we have seen, there follows immediately

$$\lim_{n \rightarrow \infty} \gamma_n = \gamma$$

The theorem is proved.

Sec. 47. Statement of the Problem of Limit Theorems for Sums

Given the double sequence

$$\left. \begin{array}{cccc} \xi_{11}, & \xi_{12}, & \dots, & \xi_{1k_1} \\ \xi_{21}, & \xi_{22}, & \dots, & \xi_{2k_2} \\ \dots & \dots & \dots & \dots \\ \xi_{n1}, & \xi_{n2}, & \dots, & \xi_{nk_n} \\ \dots & \dots & \dots & \dots \end{array} \right\} \quad (1)$$

of independent random variables in each row. We want to know to what limit distribution functions the distribution functions of the sums

$$\zeta_n = \xi_{n1} + \xi_{n2} + \dots + \xi_{nk_n}$$

can converge as n tends to infinity and what the conditions of this convergence are.

Henceforth we shall confine ourselves to the study of *elementary systems*, that is, double sequences (1) which satisfy the following conditions:

- (1) the variables ξ_{nk} have finite variances,
- (2) the variances of the sums ζ_n are bounded by a constant C not dependent on n ,

$$(3) \beta_n = \max_{1 \leq k \leq k_n} D\xi_{nk} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

The last requirement means that the effect of the separate terms on the sum becomes less and less as n increases.

The limit theorems for sums that we considered earlier quite obviously fit into this general scheme. For instance, in the theorems of DeMoivre-Laplace and Lyapunov we had the following double sequence:

$$\xi_{n1}, \xi_{n2}, \dots, \xi_{nn}$$

where

$$\xi_{nk} = \frac{\xi_k - M\xi_k}{\sqrt{\sum_{k=1}^n D\xi_k}} \quad (1 \leq k \leq n, n = 1, 2, \dots)$$

In the theorems of Bernoulli, Chebyshev and Markov concerning the law of large numbers we likewise had to do with double sequences in which the quantities

$$\xi_{nk} = \frac{\xi_k - M\xi_k}{n}$$

are taken for ξ_{nk} .

Sec. 48. Limit Theorems for Sums

Let there be an elementary system. Denote by $F_{nk}(x)$ the distribution function of the random variable ξ_{nk} and by $\bar{F}_{nk}(x)$ the distribution function of the variable $\bar{\xi}_{nk} = \xi_{nk} - M\xi_{nk}$; it is obvious that

$$F_{nk}(x) = F_{nk}(x + M\xi_{nk})$$

Theorem 1. *In order for the distribution functions of the sums*

$$\zeta_n = \xi_{n1} + \xi_{n2} + \dots + \xi_{nk_n} \quad (1)$$

to converge to a limit distribution function as $n \rightarrow \infty$, it is necessary and sufficient for infinitely divisible laws, the logarithms of the characteristic functions of which are given by the formula

$$\psi_n(t) = \sum_{k=1}^{k_n} \left\{ itM\xi_{nk} + \int (e^{itx} - 1) d\bar{F}_{nk}(x) \right\}^* \quad (2)$$

to converge to a limit law.

The limit laws for both sequences coincide.

Proof. The characteristic function of the sum (1) is

$$f_n(t) = \prod_{k=1}^{k_n} f_{nk}(t) = e^{it \sum_{k=1}^{k_n} M\xi_{nk}} \prod_{k=1}^{k_n} \bar{f}_{nk}(t) \quad (3)$$

where $f_{nk}(t)$ is the characteristic function of the random variable ξ_{nk} and $\bar{f}_{nk}(t)$ is the characteristic function of the variable $\bar{\xi}_{nk}$.

We know that for convergence of the distribution functions of the sums (1) to the limit distribution function $\Phi(x)$, it is necessary and sufficient that, as $n \rightarrow \infty$,

$$f_n(t) \rightarrow \varphi(t)$$

* If we introduce the notations

$$\gamma_n = \sum_{k=1}^{k_n} M\xi_{nk}, \quad G_n(u) = \sum_{k=1}^{k_n} \int_{-\infty}^u x^2 d\bar{F}_{nk}(x)$$

and note that $\int x d\bar{F}_{nk}(x) = 0$, then the functions $\psi_n(t)$ may be written as follows:

$$\psi_n(t) = i\gamma_n t + \int \{e^{itx} - 1 - itx\} \frac{1}{x^2} dG_n(x)$$

As we know, this means that $\psi_n(t)$ is the logarithm of the characteristic function of some infinitely divisible law.

It will be noted that the variances of ζ_n and of the infinitely divisible laws (2) coincide.

where $\varphi(t)$ is a continuous function; then $\varphi(t)$ is the characteristic function of the law $\Phi(x)$.

Let

$$\alpha_{nk} = \bar{f}_{nk}(t) - 1$$

For the variables ξ_{nk} ,

$$\alpha_n = \max_{1 \leq k \leq k_n} |\alpha_{nk}| \rightarrow 0 \quad (4)$$

uniformly in every finite interval of t . Indeed,

$$\alpha_{nk} = \int (e^{itx} - 1) d\bar{F}_{nk}(x) = \int (e^{itx} - 1 - itx) d\bar{F}_{nk}(x)$$

since

$$M\xi_{nk} = \int x d\bar{F}_{nk}(x) = 0$$

We know that for all real α

$$|e^{i\alpha} - 1 - i\alpha| \leq \frac{\alpha^2}{2}$$

Therefore,

$$|\alpha_{nk}| \leq \frac{t^2}{2} \int x^2 d\bar{F}_{nk}(x) = \frac{t^2}{2} D\xi_{nk} \quad (5)$$

From (5) and the third condition of elementariness of a system there follows (4).

From (4) we first of all conclude that for any T we can assume that for sufficiently large n and $|t| \leq T$

$$|\alpha_{nk}| < \frac{1}{2} \quad (6)$$

By virtue of this fact we can make use of the series expansion of the logarithm

$$\log \bar{f}_{nk}(t) = \log(1 + \alpha_{nk}) = \alpha_{nk} - \frac{\alpha_{nk}^2}{2} + \frac{\alpha_{nk}^3}{3} - \dots = \alpha_{nk} + r_{nk}$$

Obviously,

$$\begin{aligned} R_n &= \left| \log f_n(t) - \sum_{n=1}^{k_n} (itM\xi_{nk} + \alpha_{nk}) \right| = \\ &= \left| \sum_{n=1}^{k_n} (\log \bar{f}_{nk}(t) - \alpha_{nk}) \right| \leq \sum_{k=1}^{k_n} \sum_{n=2}^{\infty} \frac{|\alpha_{nk}|^s}{s} \leq \frac{1}{2} \sum_{k=1}^{k_n} \frac{|\alpha_{nk}|^2}{1 - |\alpha_{nk}|} \end{aligned} \quad (7)$$

Formula (5) leads to the inequality

$$R_n \leq \max_{1 \leq k \leq k_n} |\alpha_{nk}| \sum_{k=1}^{k_n} |\alpha_{nk}| \leq \frac{t^2}{2} C \max_{1 \leq k \leq k_n} |\alpha_{nk}|$$

From (4) we conclude that

$$|\log f_n(t) - \psi_n(t)| \rightarrow 0 \quad (8)$$

uniformly in every finite interval of t as n tends to infinity.

We have thus established that *in every elementary system the distribution functions of the sums ζ_n and the infinitely divisible distribution functions defined by formula (2) are asymptotic as n tends to infinity*, and so Theorem 6 is proved.

This theorem makes it possible to replace the investigation of sums (1) of random variables having, generally speaking, arbitrary distribution functions by the study of infinitely divisible laws, which, as we shall see, is in many cases extremely simple.

Theorem 2. *Every distribution law that is a limit law for the distribution functions of sums in an elementary system is infinitely divisible with finite variance and, conversely, every infinitely divisible law with finite variance is a limit law for the distribution functions of the sums of some elementary system.*

Proof. From Theorem 1 we know that the limit law for the distribution functions of sums (1) is a limit law for infinitely divisible laws and, consequently, by Theorem 3, Sec. 44, it is infinitely divisible; its variance is finite since the variances of sums are uniformly bounded by the second condition of elementariness of a system. The converse proposition that every infinitely divisible law with finite variance is a limit law for sums follows directly from the definition of infinitely divisible laws.

Theorem 3. *In order that the distribution functions of the sums (1) converge, as $n \rightarrow \infty$, to some limit distribution function and their variances converge to the variance of the limit law, it is necessary and sufficient that there exist a function $G(u)$ and a constant γ such that as n tends to infinity*

$$(1) \sum_{k=1}^{k_n} \int_{-\infty}^u x^2 d\bar{F}_{nk}(x) \rightarrow G(u)$$

at the continuity points of the functions $G(u)$,

$$(2) \sum_{k=1}^{k_n} \int x^2 d\bar{F}_{nk}(x) \rightarrow G(+\infty)$$

$$(3) \sum_{k=1}^{k_n} \int x dF_{nk}(x) \rightarrow \gamma$$

The logarithm of the characteristic function of the limit law is given by formula (1) of Sec. 45 with the function $G(u)$ and the constant γ that have just been defined.

Proof. If we introduce the notations

$$G_n(u) = \sum_{k=1}^{k_n} \int_{-\infty}^u x^2 d\bar{F}_{nk}(x)$$

and

$$\gamma_n = \sum_{k=1}^{k_n} \int x dF_{nk}(x)$$

we arrive at the conditions of the theorem of Sec. 46. That proves the theorem.

By slightly modifying the formulation of Theorem 3 we can obtain not only the conditions for the existence of the limit law, but also the conditions for convergence to every given limit law.

Theorem 4. *In order that the distribution functions of the sums (1) converge, as n tends to infinity, to a given distribution function $\Phi(x)$ and the variances of the sums converge to the variance of the limit law, it is necessary and sufficient that as n tends to infinity the following conditions be satisfied:*

$$(1) \sum_{k=1}^{k_n} \int_{-\infty}^u x^2 d\bar{F}_{nk}(x) \rightarrow G(u)$$

at the continuity points of the function $G(u)$

$$(2) \sum_{k=1}^{k_n} \int x^2 d\bar{F}_{nk}(x) \rightarrow G(\infty)$$

$$(3) \sum_{k=1}^{k_n} \int x dF_{nk}(x) \rightarrow \gamma$$

where the function $G(u)$ and the constant γ are given by formula (1) of Sec. 45 for the function $\Phi(x)$.

Sec. 49. Conditions for Convergence to the Normal and Poisson Laws

We shall apply the results of Sec. 48 in order to derive the conditions for the convergence of the distribution functions of sums to the normal and Poisson laws.

Theorem 1. *Given an elementary system of independent random variables. For the distribution functions of the sums*

$$\zeta_n = \xi_{n1} + \xi_{n2} + \dots + \xi_{nk_n} \quad (1)$$

to converge, as n tends to infinity, to the law

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx$$

it is necessary and sufficient that, as n tends to infinity, the following conditions be satisfied:

$$\begin{aligned} (1) \quad & \sum_{k=1}^{k_n} \int x dF_{nk}(x) \rightarrow 0 \\ (2) \quad & \sum_{k=1}^{k_n} \int_{|x| > \tau} x^2 d\bar{F}_{nk}(x) \rightarrow 0 \\ (3) \quad & \sum_{k=1}^{k_n} \int_{|x| < \tau} x^2 d\bar{F}_{nk}(x) \rightarrow 1 \end{aligned}$$

where τ is any positive constant.

Proof. From Theorem 4, Sec. 48, it follows that the desired conditions consist in satisfaction of the following relations, as n tends to infinity,

$$\begin{aligned} & \sum_{k=1}^{k_n} \int x dF_{nk}(x) \rightarrow 0 \\ & \sum_{k=1}^{k_n} \int_{-\infty}^u x^2 d\bar{F}_{nk}(x) \rightarrow \begin{cases} 0 & \text{for } u < 0 \\ 1 & \text{for } u > 0 \end{cases} \\ & \sum_{k=1}^{k_n} \int x^2 d\bar{F}_{nk}(x) \rightarrow 1 \end{aligned}$$

The first one coincides with the first condition of the theorem, and it is obvious that the two others are equivalent to the second and third conditions of the theorem.

This theorem takes on an especially simple form if the elementary system under consideration is normalized beforehand by the conditions

$$\left. \begin{aligned} \sum_{k=1}^{k_n} \int x^2 dF_{nk}(x) &= 1 \\ \int x dF_{nk}(x) &= 0 \quad (1 \leq k \leq k_n, n = 1, 2, \dots) \end{aligned} \right\} \quad (2)$$

Theorem 2. *If an elementary system is normalized by the relations (2), then for convergence of the distribution functions of the sums (1) to the normal law it is necessary and sufficient that for all $\tau > 0$, as n tends to infinity,*

$$\sum_{k=1}^{k_n} \int_{|x| > \tau} x^2 dF_{nk}(x) \rightarrow 0 \quad (3)$$

The *proof* of the theorem is obvious.

The requirement (3) bears the name of *Lindeberg's condition* because Lindeberg, in 1923, proved its sufficiency for convergence of the distribution functions of sums to the normal law. In 1935 W. Feller proved the necessity of this condition.

To illustrate the use of the general theorems of the preceding section we consider the convergence of the distribution functions of elementary systems to the Poisson law:

$$P(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \sum_{0 \leq k < x} e^{-\lambda} \frac{\lambda^k}{k!} & \text{for } x > 0 \end{cases} \quad (4)$$

If ξ is a random variable distributed in accordance with the law (4), then, as we know, $M\xi = D\xi = \lambda$.

We confine ourselves to elementary systems for which

$$\left. \begin{aligned} \sum_{k=1}^{k_n} M\xi_{nk} &\rightarrow \lambda \\ \sum_{k=1}^{k_n} D\xi_{nk} &\rightarrow \lambda \end{aligned} \right\} \quad (5)$$

Theorem 3. *Let there be given an elementary system that obeys the conditions (5). The distribution functions of the sums*

$$\zeta_n = \xi_{n1} + \xi_{n2} + \dots + \xi_{nk_n}$$

converge to the law (4) if and only if for any $\tau > 0$

$$\sum_{k=1}^{k_n} \int_{|x-1| > \tau} x^2 dF_{nk}(x + M\xi_{nk}) \rightarrow 0 \quad (n \rightarrow \infty)$$

We leave the proof of this theorem to the reader.

In Sec. 15 we proved the Poisson theorem. It will readily be seen that when $np_n = \lambda$ it is a special case of the proposition that has just been proved. Indeed, let ξ_{nk} ($1 \leq k \leq n$) be a random variable that

takes on values 0 or 1 depending on the occurrence or nonoccurrence, in the k th trial of the n th series of trials, of event A that we are observing. Here

$$\mathbf{P} \{ \xi_{nk} = 1 \} = \frac{\lambda}{n} \text{ and } \mathbf{P} \{ \xi_{nk} = 0 \} = 1 - \frac{\lambda}{n}$$

Obviously, the sum

$$\mu_n = \xi_{n1} + \xi_{n2} + \dots + \xi_{nn}$$

is the number of occurrences of the event A in the n th series of trials.

According to the Poisson theorem, the distribution functions of the variables μ_n reduce to the Poisson law (5) as $n \rightarrow \infty$. This result also follows from the theorem that has just been formulated, since in the given case all its requirements are satisfied.

The general theorems concerning the approach of the distribution functions of the sums (1) to some infinitely divisible distribution functions, proved under broader assumptions than ours, also permit obtaining the necessary and sufficient condition for the law of large numbers (in the case of independent summands). See the earlier mentioned monograph of B. V. Gnedenko and A. N. Kolmogorov.

EXERCISES

1. Prove that the distributions of
 - (a) Pascal (Exercise 1 (a) of Chapter 5),
 - (b) Polya (Exercise 1 (b) of Chapter 5),
 - (c) Cauchy (Example 5, Sec. 24)
 are infinitely divisible.

2. Prove that a random variable with density function

$$p(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0 \end{cases}$$

where $\alpha > 0$, $\beta > 0$ are constants, is infinitely divisible.

Note. From this it follows, in particular, that the Maxwell distribution and the chi-square distribution are infinitely divisible for any value of n .

3. Prove that, no matter what the constants $\alpha > 0$ and $\beta > 0$,

$$\varphi(t) = \left(1 + \frac{t^2}{\beta^2} \right)^{-\alpha}$$

is an infinitely divisible characteristic function.

Note. From this it follows, in particular, that the Laplace distribution (Exercise 6 of Chapter 5) is infinitely divisible.

4. Find the function $G(x)$ and the parameter γ in Kolmogorov's formula for the logarithm of an infinitely divisible characteristic function for:

- (a) the distribution in Example 2,
- (b) the Laplace distribution.

5. Prove that if the sum of two independent infinitely divisible random variables is distributed according to

- (a) the Poisson law,
- (b) the normal law,

then every summand is Poisson distributed in case (a) and normally distributed in case (b).

6. Find the conditions under which the distribution functions of sums of random variables constituting an elementary system converge to:

- (a) the distribution of Example 2,
- (b) the Laplace distribution.

CHAPTER 10

The Theory of Stochastic Processes

Sec. 50. Introductory Remarks

Refinements in the statistics of physics and in a number of branches of technology have confronted probability theory with a large number of new problems that do not fit into the framework of the classical theory. Whereas physics and technology were interested in studying *processes*, that is, phenomena that take place in time, the theory of probability did not have either general procedures or elaborated particular schemes for solving problems that arise in the study of such phenomena. It was insistently necessary to develop a general theory of random processes, a theory which would study random variables dependent on one or several continuously varying parameters.

Let us examine a number of problems that will illustrate the necessity of constructing a theory of random processes.

Suppose our purpose is to trace the movement of some molecule of a gas or liquid. At random instants the molecule collides with other molecules and thus alters its velocity and position. The state of the molecule is thus subjected to random changes at every instant of time. Many physical phenomena require the ability to compute the probabilities that a definite number of molecules will be able, within a given time interval, to cover certain distances. For example, if two gases or two liquids are brought into contact, a mutual penetration of the molecules of one into those of the other sets in: diffusion occurs. How fast does the diffusion process develop, according to what kind of laws, and when does the resulting mixture become practically homogeneous? All these and many other questions are answered by the statistical theory of diffusion, which is based on the theory of *random (stochastic or probabilistic) processes*. It is obvious that a similar problem arises in chemistry when studying the process of a chemical reaction. What portion of the molecules has already ente-

red into the reaction, how is the reaction proceeding in time, and when, for all practical purposes, will it come to an end?

An extremely important range of phenomena take place in accordance with the principle of radioactive disintegration. This consists in the atoms of a radioactive substance decaying into the atoms of another element. The decay of every atom occurs instantaneously, like an explosion, with release of a certain amount of energy. Numerous observations have shown that the decay of different atoms takes place at randomly chosen instants of time, as far as the observer is concerned. And the times of occurrence are independent of one another in the meaning of probability theory. For studies of the process of radioactive decay it is essential to determine the probability that within a certain time interval a certain quantity of atoms will disintegrate. Formally speaking, other phenomena proceed in exactly the same fashion if we confine ourselves to elucidating the mathematical picture of the phenomenon. Such are the number of calls at a telephone exchange during a given time interval (the load or traffic at the telephone exchange), the breakage of thread on a ring spinning frame (type of spinning loom) or changes in the number of particles that are in Brownian motion and that are located in a given region of space at a given instant of time. In this chapter we give a simple solution to the mathematical problems that such phenomena lead to.

We have considered some simple problems of a practical nature concerned with concrete stochastic processes (see Chapter 1, Sec. 10, Examples 2, 3, 4 and Exercises 21 and 22 of Chapter 1, and 15 and 16 of Chapter 2).

To what has been said let us add that in the introduction it was stated that the first examples of probabilistic processes were considered at the beginning of this century by a number of outstanding physicists. We shall now give a brief description of how by proceeding from the extremely schematic random walk problem, they were able to obtain the differential equation of diffusion theory. The line of reasoning is as follows: let a particle be subject, at times $k\tau$ ($k=1, 2, \dots$), to independent random impacts as a result of which it is displaced each time to the right by an amount h with probability p and to the left by an amount h with probability $q=1-p$. Denote by $f(x, t)$ the probability that the moving particle, after starting from the point $x=0$ at time $t=0$, will, as a result of n impacts, reach the position x at time t ($t=n\tau$). It is clear that in the case of an even number of impacts the quantity x will only equal an even number of steps h and, for an odd number n , the number of steps h will be odd. If by m we denote the number of steps taken by the particle to the right ($n-m$, respectively, the number of steps to the left), then, according to the Bernoulli formula,

$$f(x, t) = C_n^m p^m q^{n-m}$$

It is clear that the quantities m, n, x, τ are connected by the equation

$$m - (n - m) = \frac{x}{h}$$

Direct calculation immediately shows that $f(x, t)$ satisfies the following difference equation:

$$f(x, t + \tau) = pf(x - h, t) + qf(x + h, t) \quad (1)$$

and the initial conditions

$$f(0, 0) = 1, \quad f(x, 0) = 0 \quad \text{for } x \neq 0$$

Let us see how the difference equation changes if we let both h and τ tend to zero. The physical nature of the problem will compel us to impose certain restrictions on h and τ . By the same reasoning, the quantities p and q cannot be taken arbitrarily. Non-observance of certain conditions discussed below can result in the particle, within a finite interval of time, going off to infinity with probability one. To escape this possibility, we impose the following requirements: when n tends to infinity,

$$x = nh, \quad t = n\tau, \quad \frac{h^2}{\tau} \rightarrow 2D, \quad \frac{p - q}{h} \rightarrow \frac{c}{D} \quad (2)$$

where c and D are certain constants, the former called the *drift*, the latter, the *diffusion coefficient*.

Subtracting $f(x, t)$ from both sides of (1), we get

$$f(x, t + \tau) - f(x, t) = p[f(x - h, t) - f(x, t)] + q[f(x + h, t) - f(x, t)] \quad (3)$$

Suppose that $f(x, t)$ is differentiable with respect to t and twice differentiable with respect to x . Then

$$\begin{aligned} f(x, t + \tau) - f(x, t) &= \tau \frac{\partial f(x, t)}{\partial t} + o(\tau) \\ f(x - h, t) - f(x, t) &= -h \frac{\partial f(x, t)}{\partial x} + \frac{1}{2} h^2 \frac{\partial^2 f(x, t)}{\partial x^2} + o(h^2) \\ f(x + h, t) - f(x, t) &= +h \frac{\partial f(x, t)}{\partial x} + \frac{1}{2} h^2 \frac{\partial^2 f(x, t)}{\partial x^2} + o(h^2) \end{aligned}$$

Substituting these equations into (3) we get

$$\tau \frac{\partial f(x, t)}{\partial t} + o(\tau) = -(p - q)h \frac{\partial f(x, t)}{\partial x} + \frac{h^2}{2} \frac{\partial^2 f(x, t)}{\partial x^2} + o(h^2)$$

And from this, by virtue of the relations (2), we find that in the limit

$$\frac{\partial f(x, t)}{\partial t} = -2c \frac{\partial f(x, t)}{\partial x} + D \frac{\partial^2 f(x, t)}{\partial x^2}$$

We have arrived at an equation that in diffusion theory is called the *Fokker-Planck equation*.

It is interesting to note that this rather artificial statement of the problem has yielded a physically meaningful result which reflects very well the true picture of diffusion. Later on we will derive the general equations obeyed by distributions for stochastic processes under extremely broad assumptions concerning the nature of their development.

The general theory of stochastic processes originated in the fundamental works of the Soviet mathematicians A. N. Kolmogorov and A. Ya. Khinchin at the beginning of the 1930s. Kolmogorov, in a paper entitled "On Analytical Methods in Probability Theory", gave a systematic and rigorous construction of the fundamentals of the theory of *stochastic processes without aftereffect* or, as it is customary to say, *Markov processes*. In a number of works, Khinchin created the principles of the theory of so-called *stationary processes*.

Before subjecting natural or technical processes to a mathematical study, they first have to be made schematic. The reason for this lies in the fact that mathematical analysis is applicable to the investigation of a process of variation of some system only if it is assumed that every possible state of the system is fully defined by means of some definite mathematical apparatus. Quite naturally, such a mathematically defined system is not actuality itself, but only a scheme suitable for its description. That, for instance, is what we encounter in mechanics when it is assumed that the real motions of a system of mass points can be completely described for any instant of time by indicating the instant and its state at any preceding time t_0 . In other words, the scheme used in theoretical mechanics for describing motion consists in the following: it is taken that for any time t the state of a system y is fully determined by its state x at any preceding time t_0 . Here, the state of the system in mechanics is understood to be the specification of positions and velocities of the points of the material system.

Outside classical mechanics, actually throughout the whole of modern physics, one has to do with a far more complicated situation when a knowledge of the state of a system at some time t_0 no longer uniquely determines the state of the system at subsequent times but only determines the probability that the system will be in one of the states of a certain set of states of the system. If by x we denote the state of the system at time t_0 and by E a certain collection of states of the system, then for the processes just described there is defined the probability

$$P\{t_0, x; t, E\}$$

that the system which at time t_0 is in the state x will at time t pass into one of the states of the set E .

If any additional knowledge of the states of the system at times $t < t_0$ does not alter the probability, then it is natural to call this class of stochastic processes that we have isolated *processes without aftereffect*, or, by analogy with Markov chains, *Markov processes*.

The general concept of a stochastic process that is based on the earlier presented axiomatics may be introduced as follows: Let U be a set of elementary events and t a continuous parameter. A *stochastic process* is defined as the function of two arguments:

$$\xi(t) = \varphi(e, t) \quad (e \in U)$$

For every value of the parameter t , the function $\varphi(e, t)$ is a function of e only and, consequently, is a random variable. For every fixed value of the argument e (that is, for every elementary event) $\varphi(e, t)$ depends only on t and is thus simply a function of one real argument. Every such function is called a *realization* of the stochastic process $\xi(t)$. We may regard a stochastic process either as a collection of random variables $\xi(t)$ that depend on the parameter t , or as a collection of the realizations of the process $\xi(t)$. Naturally, to define a process it is necessary to specify a probability measure in the function space of its realizations.

The present chapter will be devoted, in its entirety, to a study of processes without aftereffect and of stationary processes.

Sec. 51. The Poisson Process

Before beginning an exposition of some of the general results that have now become classical, we will make a detailed study of one example of a stochastic process without aftereffect that plays an important role both in theory and in a diversity of applications.

Suppose a certain event occurs at random times. We are interested in the number of occurrences of the event during the time interval from 0 to t . Denote that number by $\xi(t)$. Relative to the process of occurrence of the event we will presume that it is (1) stationary; (2) without aftereffect, and (3) ordinary. The following meaning is attached to these assumptions.

Stationarity signifies that for any group of a finite number of nonoverlapping time intervals the probability of occurrence of a definite number of events during the course of each one of them depends on the numbers only and on the duration of the time intervals, but is not changed by an identical shift in all the time intervals. In particular, the probability of occurrence of k demands (events) during the time interval from T to $T+t$ is independent of T and is a function only of k and t .

The *absence of aftereffect* means that the probability of occurrence of k events during the time interval $(T, T+t)$ does not depend on how many times the events occurred previously or how they occurred.

This assumption means that the conditional probability of occurrence of k events during the time interval $(T, T+t)$ under any assumption of the occurrence of events prior to time T coincides with the unconditional probability. In particular the absence of aftereffect signifies mutual independence of the occurrence of any number of events during nonoverlapping intervals of time.

Ordinariness expresses the requirement of practical impossibility of the occurrence of two or several events during a small time interval Δt . Denote by $P_{>1}(\Delta t)$ the probability of occurrence of more than one event in the time interval Δt . Then the condition of ordinariness, precisely expressed, consists in the following:

$$P_{>1}(\Delta t) = o(\Delta t)$$

Our immediate problem will then be to determine the probability $P_k(t)$ that during an interval of duration t there will occur k events. By the assumptions made, these probabilities do not depend on the location of the time interval. With this purpose in mind, we find that for small Δt the following equation holds:

$$P_1(\Delta t) = \lambda \Delta t + o(\Delta t)$$

where λ is a constant.

Indeed, consider a time interval of duration 1 and denote by p the probability that no event will occur within this period. Partition the time interval into n nonoverlapping equal parts. By virtue of stationarity and the absence of aftereffect we have

$$p = \left[P_0\left(\frac{1}{n}\right) \right]^n$$

and so

$$P_0\left(\frac{1}{n}\right) = p^{\frac{1}{n}}$$

From this, for any integral k

$$P_0\left(\frac{k}{n}\right) = p^{\frac{k}{n}}$$

Now let t be some nonnegative number. For any n it is possible to find a k such that

$$\frac{k-1}{n} \leq t < \frac{k}{n}$$

Since the probability $P_0(t)$ is a decreasing function of time,

$$P_0\left(\frac{k-1}{n}\right) \geq P_0(t) \geq P_0\left(\frac{k}{n}\right)$$

Thus, $P_0(t)$ satisfies the inequalities

$$p^{\frac{k-1}{n}} \geq P_0(t) \geq p^{\frac{k}{n}}$$

Now let k and n tend to infinity so that

$$\lim_{n \rightarrow \infty} \frac{k}{n} = t$$

From the foregoing it is clear that

$$P_0(t) = p^t$$

Since $P_0(t)$, being a probability, satisfies the inequalities

$$0 \leq P_0(t) \leq 1$$

three cases are possible: (1) $p=0$; (2) $p=1$; (3) $0 < p < 1$. The first two cases are of little interest. In the first we have the equation $P_0(t)=0$ for any t and, hence, the probability is one that at least one event will occur during a time interval of any length. In other words, infinitely many events will occur with probability 1 during a time interval of arbitrary duration. In the second case, $P_0(t)=1$, and, consequently, events do not occur. Only the third case is of interest; here put $p=e^{-\lambda}$ where λ is some positive number ($\lambda=-\ln p$).

Summarizing, then, from the assumptions of stationarity and absence of aftereffect, we have found that for any $t \geq 0$ (we have not yet made use of the assumption of ordinariness)

$$P_0(t) = e^{-\lambda t} \quad (1)$$

Clearly the equation

$$P_0(t) + P_1(t) + P_{>1}(t) = 1$$

holds for any value of t .

It follows from the foregoing that for small t

$$P_0(t) = 1 - \lambda t + o(t)$$

Hence for small t

$$P_1(t) = \lambda t + o(t) \quad (2)$$

Now we can derive the formulas for the probabilities $P_k(t)$ for $k \geq 1$. For this purpose, we determine the probability that during time $t + \Delta t$ an event will occur exactly k times. This may occur in $k+1$ different ways, namely:

(1) during time interval t all k events will occur, and none will occur during time Δt ;

(2) during an interval of length t there will occur $k-1$ events, and during time Δt , only one; ...

($k+1$) during the time interval t the event will not occur once, but during Δt it will occur k times.

By the formula of total probability,

$$P_k(t + \Delta t) = \sum_{j=0}^k P_j(t) P_{k-j}(\Delta t)$$

(here both the condition of stationarity and absence of aftereffect have been taken into account). Put

$$R_k = \sum_{j=0}^{k-2} P_j(t) P_{k-j}(\Delta t)$$

It is obvious that

$$R_k \leq \sum_{j=0}^{k-2} P_{k-j}(\Delta t) = \sum_{s=2}^k P_s(\Delta t) \leq \sum_{s=2}^{\infty} P_s(\Delta t) = P_{>1}(\Delta t) = o(\Delta t)$$

according to the condition of ordinariness.

Thus,

$$P_k(t + \Delta t) = P_k(t) P_0(\Delta t) + P_{k-1}(t) P_1(\Delta t) + o(\Delta t)$$

But from what has been proved,

$$P_0(\Delta t) = e^{-\lambda \Delta t} = 1 - \lambda \Delta t + o(\Delta t)$$

Furthermore, by (2),

$$P_1(\Delta t) = \lambda \Delta t + o(\Delta t)$$

and so

$$P_k(t + \Delta t) = (1 - \lambda \Delta t) P_k(t) + \lambda \Delta t P_{k-1}(t) + o(\Delta t)$$

From this we have

$$\frac{P_k(t + \Delta t) - P_k(t)}{\Delta t} = -\lambda P_k(t) + \lambda P_{k-1}(t) + o(1)$$

Since as $\Delta t \rightarrow 0$ the limit of the right-hand side of the equation exists, the left-hand side also has a limit. As a result we get the equation

$$\frac{dP_k(t)}{dt} = -\lambda P_k(t) + \lambda P_{k-1}(t) \quad (3)$$

for the determination of $P_k(t)$. Obviously the requirement of ordinariness and the expression for $P_0(t)$ that we found lead to the following initial conditions:

$$P_0(0) = 1; P_k(0) = 0 \text{ for } k \geq 1 \quad (4)$$

It is easy to solve Equations (3) by making the substitution

$$P_k(t) = e^{-\lambda t} v_k(t) \quad (5)$$

where $v_k(t)$ is the new desired function. Note that, by (1), $v_0(t) = 1$. The relations (4) lead to the following initial conditions:

$$v_0(0) = 1 \text{ and } v_k(0) = 0 \text{ for } k \geq 1 \quad (6)$$

Substitution of (5) into (3) gives us

$$\frac{dv_k(t)}{dt} = \lambda v_{k-1}(t) \quad (7)$$

In particular,

$$\frac{dv_1(t)}{dt} = \lambda \quad (7')$$

Solution of Equations (7') and (7) in succession brings us (taking into account the initial conditions) to the equation

$$v_1(t) = \lambda t, \quad v_2(t) = \frac{(\lambda t)^2}{2!}, \quad v_3(t) = \frac{(\lambda t)^3}{3!}$$

and, generally,

$$v_k(t) = \frac{(\lambda t)^k}{k!}$$

We thus get

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (8)$$

for any $k \geq 0$. Our problem is solved.

The requirements satisfied by the process of occurrence of the event are fulfilled to a high degree of accuracy in numerous natural phenomena and technological processes. We have instances such as the number of spontaneously disintegrating atoms of a radioactive substance during a given time interval and the number of cosmic-ray particles impinging on a definite area during time t . If we have to do with some kind of complicated electronic system consisting of a large number of elements, each of which can break down with a small probability during unit time and independently of the states of the other elements, then the number of elements failing in the time interval $(0, t)$ is a stochastic process. In many cases, this process is well described by the Poisson process. There is literally no limit to the number of such examples.

Here we shall examine two simple properties of Poisson processes.

The time interval between two successive occurrences of some event that interests us is a random variable, which we denote by τ . Find the probability distribution of τ . Since it is obvious that the event $\tau > t$ is equivalent to no event occurring in the time interval t that we record,

$$P\{\tau > t\} = e^{-\lambda t}$$

The desired distribution function is thus given by the formula

$$P\{\tau < t\} = 1 - e^{-\lambda t} \quad (9)$$

This result may be interpreted physically in many ways. For example, we can regard it as a time distribution of the free motion of a molecule or as the time distribution between two failures of elements in a complex electronic system.

Suppose we know that during an interval of length t there occur n ($n > 0$) events of our process. The question is: under that condition, how are the occurrences of these events distributed within the given time interval? It turns out that the conditional distribution of times of occurrence of the events is uniform in this time interval. Moreover, the instants of occurrence of all n events are mutually independent.

Denote by B the event which consists in the occurrence of n events of the process during the time interval $(0, t)$. We know from the foregoing that the probability of B is

$$P(B) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Insofar as the events have already come within the time interval $(0, t)$, we can individualize them and concentrate our attention on one of them. Denote by A the event which consists in the fact that the event which interests us occurred in the interval (a, b) belonging to $(0, t)$. Our problem is to determine the probability $P(A/B)$. By the multiplication theorem,

$$P(A/B) = \frac{P(AB)}{P(B)}$$

We have to determine the probability of the joint occurrence of events A and B . For this purpose consider the events C_{sr} , which consist in the fact that: (a) some kind of s events of the process will fall in $(0, a)$ but the one that interests us will not be among them; (b) r events of the process including the one we are interested in will occur in the interval (a, b) ; (c) the remaining $n-r-s$ events will fall in the interval (b, t) . Obviously, the events C_{sr} are incompatible for distinct pairs (s, r) and for this reason

$$AB = \sum_{s=0}^{n-1} \sum_{r=1}^{n-s} C_{sr}$$

By the assumption concerning the absence of aftereffect, the probability of obtaining s events of the process in $(0, a)$, r events in (a, b) and $n-s-r$ in the interval (b, t) is

$$\frac{(\lambda a)^s}{s!} e^{-\lambda a} \frac{[\lambda (b-a)]^r}{r!} e^{-\lambda (b-a)} \frac{[\lambda (t-b)]^{n-r-s}}{(n-s-r)!} e^{-\lambda (t-b)} \quad (10)$$

This expression, however, is different from the probability of the event C_{sr} , for we did not take into account the necessity of our particular event occurring in the interval (a, b) . In order to take

this circumstance into account we also have to multiply (10) by the probability that the event which interests us will fall in the interval (a, b) . This probability is equal to the ratio of the number of ways of selecting $r-1$ elements from $n-1$ to the number of ways of choosing r elements out of n ; that is, it is equal to

$$\frac{C_{n-1}^{r-1}}{C_n^r} = \frac{r}{n}$$

Thus,

$$P(AB) = e^{-\lambda t} \lambda^n \sum_{s=0}^{n-1} \sum_{r=1}^{n-s} \frac{r}{n} \frac{a^s}{s!} \frac{(b-a)^r}{r!} \frac{(t-b)^{n-r-s}}{(n-r-s)!}$$

Simple algebra leads us to the equation

$$P(BA) = \frac{b-a}{t} \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

Collecting the computations, we find

$$P(A/B) = \frac{b-a}{t} \quad (11)$$

This equation proves the formulated result.

We note that the theory developed in this section may be applied not only on the assumption that the parameter t plays the role of time. With this in mind, we add a further example.

Example. Points are dispersed in space in accordance with the following requirements:

(1) the probability that k points will be found in the region G depends solely on the volume v of the region but is independent both of its shape and its position in space; denote this probability by the symbol $p_k(v)$;

(2) the numbers of points that fall in nonoverlapping regions are independent random variables;

$$(3) \quad \sum_{k=2}^{\infty} p_k(\Delta v) = o(\Delta v)$$

The conditions that have been imposed are nothing other than conditions of stationarity, absence of aftereffect, and ordinariness. And so

$$p_n(v) = \frac{(av)^n}{n!} e^{-av}$$

If minute particles of some substance are suspended in a liquid, then under the impacts of surrounding molecules these particles will be in a state of constant chaotic motion (Brownian motion). As a result, at each instant of time we have a random distribution of particles in space, which is exactly what we have been speaking about.

According to the theory of this example, we must consider that the distribution of particles entering some definite region will be subject to the Poisson law.

Table 12 compares the results of experiments with particles of gold suspended in water and computations by Poisson's law (taken from a paper by Smolukhovsky).

TABLE 12

Number of particles	Number of observed cases	Frequency, $\frac{m}{518}$	$\frac{\lambda^n e^{-\lambda}}{n!}$	Number of cases computed
0	112	0.216	0.213	110
1	168	0.325	0.328	173
2	130	0.251	0.253	131
3	69	0.133	0.130	67
4	32	0.062	0.050	26
5	5	0.010	0.016	8
6	1	0.002	0.004	2
7	1	0.002	0.001	1

The constant $\lambda = av$ that defines the Poisson law is chosen equal to the arithmetic mean of the observed number of particles, that is

$$\lambda = \frac{0 \times 112 + 1 \times 168 + 2 \times 130 + 3 \times 69 + 4 \times 32 + 5 \times 5 + 6 \times 1 + 7 \times 1}{518} \approx 1.54$$

Sec. 52. Conditional Distribution Functions and Bayes' Formula

For the further development of the theory we have to generalize the concept of conditional probability that was introduced in the first chapter to the case of an infinite number of possible conditions. In particular, we have to introduce the concept of a conditional distribution function with respect to some random variable.

Consider a certain event B and a random variable ξ with a distribution function $F(x)$. Denote by $A_{\alpha\beta}$ the event that

$$x - \alpha \leq \xi < x + \beta$$

By virtue of the definitions of Chapter 1,

$$\mathbf{P}\{BA_{\alpha\beta}\} = \mathbf{P}\{A_{\alpha\beta}\} \mathbf{P}\{B/A_{\alpha\beta}\} = [F(x + \beta) - F(x - \alpha)] \mathbf{P}\{B/A_{\alpha\beta}\}$$

which gives

$$\mathbf{P}\{B/A_{\alpha\beta}\} = \frac{\mathbf{P}\{BA_{\alpha\beta}\}}{F(x + \beta) - F(x - \alpha)}$$

The limit

$$\lim_{\alpha, \beta \rightarrow 0} \frac{P\{BA_{\alpha\beta}\}}{F(x+\beta) - F(x-\alpha)}$$

if it exists* is called the *conditional probability of event B provided that $\xi=x$* , and is denoted by the symbol $P\{B/x\}$. It is obvious that for fixed x , $P\{B/x\}$ will be a finitely additive function of B defined on some field of events.

Under certain conditions, which are practically always fulfilled, $P\{B/x\}$ will have all the properties of ordinary probability satisfying the Axioms 1 to 3 of Sec. 8.

If η is a random variable and B denotes the event that $\eta < y$, then the function $\Phi(y/x) = P\{\eta < y/x\}$, which, as it is easy to see, will be a distribution function, is called the *conditional distribution function of the variable η provided that $\xi=x$* .

It is obvious that if $F(x, y)$ is the distribution function of a pair of random variables ξ and η , then

$$\Phi(y/x) = \lim_{\alpha, \beta \rightarrow 0} \frac{F(x+\beta, y) - F(x-\alpha, y)}{F(x+\beta, \infty) - F(x-\alpha, \infty)}$$

provided that this limit exists.

If the function $P\{B/x\}$ is integrable with respect to $F(x)$, we have the *formula of total probability*:

$$P\{B\} = \int P\{B/x\} dF(x)$$

To prove this formula we divide the interval of variation of the variable ξ by the points x_i ($i=0, \pm 1, \pm 2, \dots$) into the subintervals $x_i \leq \xi < x_{i+1}$. Denote by A_i the event $x_i \leq \xi < x_{i+1}$. By virtue of the extended axiom of addition we have

$$P\{B\} = \sum_{i=-\infty}^{\infty} P\{BA_i\} = \sum_{i=-\infty}^{\infty} P\{B/A_i\} [F(x_{i+1}) - F(x_i)]$$

We now partition the subintervals (x_i, x_{i+1}) into still smaller subintervals so that the maximal length of the subintervals thus obtained should tend to zero. From this, by the definition of conditional probability and by the Stieltjes integral we get

$$P\{B\} = \int P\{B/x\} dF(x)$$

In particular,

$$\Phi(y) = P\{\eta < y\} = \int \Phi(y/x) dF(x) \quad (1)$$

* This limit exists for almost all values of x , in the sense of the measure defined by the function $F(x)$.

If there exists a probability density function of the variable η , then

$$\varphi(y) = \int \varphi(y/x) dF(x) \quad (1')$$

where $\varphi(y/x)$ is the *conditional density function* of the variable η .

Example. To illustrate the use of formula (1), let us consider the following problem in the theory of gunfire. Errors of two kinds are involved in firing at a target: (1) errors in defining the position of the target and (2) errors of fire due to a large number of diverse causes (variations in the size of the charge in the shell, irregularities of machining of the casing of the shell, errors in sighting, slight fluctuations in atmospheric conditions, etc.). Errors of the second kind are called technical dispersion.

A total of n independent shots are fired at one defined position of the target. It is required to determine the probability of one hit.

For the sake of simplicity, we confine ourselves to a consideration of a one-dimensional target of size 2α and the shell will be assumed a point. Denote by $f(x)$ the density function of the position of the target and by $\varphi_i(x)$ the density function for the points of impact of the i th shell.

If the centre of the target lies at point z , then the probability of hitting the target in the i th shot is equal to the probability of falling in the interval $(z-\alpha, z+\alpha)$, that is, it is equal* to

$$\int_{z-\alpha}^{z+\alpha} \varphi_i(x) dx$$

The conditional probability of a miss in the i th shot provided the centre of the target lies at point z is

$$1 - \int_{z-\alpha}^{z+\alpha} \varphi_i(x) dx$$

The conditional probability of a miss for all n shots (given the same condition) is equal to

$$\prod_{i=1}^n \left(1 - \int_{z-\alpha}^{z+\alpha} \varphi_i(x) dx \right)$$

Whence we conclude that the probability of at least one hit, given that the centre of the target is at z , is equal to

$$1 - \prod_{i=1}^n \left(1 - \int_{z-\alpha}^{z+\alpha} \varphi_i(x) dx \right)$$

* We assume here that the determination of the target position and the technical dispersion are independent.

The unconditional probability of at least one hit (by formula (1)) is thus

$$P = \int f(z) \left[1 - \prod_{i=1}^n \left(1 - \int_{z-\alpha}^{z+\alpha} \varphi_i(x) dx \right) \right] dz$$

If the firing conditions do not change from shot to shot, then $\varphi_i(x) = \varphi(x)$ ($i = 1, 2, \dots, n$) and consequently

$$P = \int f(z) \left[1 - \left(1 - \int_{z-\alpha}^{z+\alpha} \varphi(x) dx \right)^n \right] dz$$

As before, let A_i denote the event $x_i \leq \xi < x_{i+1}$. According to the classical theorem of Bayes

$$\mathbf{P}\{A_i/B\} = \frac{\mathbf{P}\{A_i\} \mathbf{P}\{B/A_i\}}{\mathbf{P}\{B\}}$$

If $F(x) = \mathbf{P}\{\xi < x\}$ and $\mathbf{P}\{\xi < x/B\}$ have continuous derivatives with respect to x , then, using the Lagrange theorem, we get

$$\mathbf{P}\{A_i/B\} = p_{\xi}(\bar{x}_i/B) (x_{i+1} - x_i) = \frac{F'(\bar{x}_i) \mathbf{P}\{B/A_i\}}{\mathbf{P}\{B\}} (x_{i+1} - x_i)$$

where $x_i < \bar{x}_i < x_{i+1}$, $x_i < \bar{x}'_i < x_{i+1}$. In the limit, when $x_i \rightarrow x$; $x_{i+1} \rightarrow x$, we get

$$p_{\xi}(x/B) = \frac{\rho(x) \mathbf{P}\{B/x\}}{\mathbf{P}\{B\}}$$

or

$$p_{\xi}(x/B) = \frac{\rho(x) \mathbf{P}\{B/x\}}{\int \mathbf{P}\{B/x\} \rho(x) dx} \quad (2)$$

It is natural to call this equality *Bayes' formula*.

Now let the event B consist in the fact that some random variable η takes on a value between $y - \alpha$ and $y + \beta$ and let the conditional distribution function $\Phi(y/x)$ of the variable η have, for every x , the continuous density $p_{\eta}(y/x)$. Then, as follows from Equation (2), if $\frac{1}{\beta + \alpha} \mathbf{P}\{B/x\}$ tends to $p_{\eta}(y/x)$ uniformly in x as α and β tend to zero, then the following equality holds:

$$p_{\xi}(x/y) = \frac{\rho(x) p_{\eta}(y/x)}{\int p_{\eta}(y/x) \rho(x) dx}$$

We shall make extensive use of this formula in the next chapter.

Sec. 53. Generalized Markov Equation

We now take up the study of stochastic processes without aftereffect confining ourselves only to the most *elementary* problems. We assume, in particular, that the set of possible states of the system is the set of real numbers. Thus, a *stochastic process* is a set of random variables $\xi(t)$ dependent on a single real parameter t . We shall call parameter t the time and speak of the state of the system at one or another instant of time.

We will obtain a complete probabilistic characteristic of a *process without aftereffect* by specifying the function $F(t, x; \tau, y)$, which is equal to the probability that at time τ the random variable $\xi(\tau)$ will take on a value less than y if it is known that at time t ($t < \tau$) we had the equality $\xi(t) = x$. Additional knowledge about the states of the system at times prior to t does not alter the function $F(t, x; \tau, y)$ for processes without aftereffect.

Let us now note some conditions that the function $F(t, x; \tau, y)$ must satisfy. First of all, since it is a distribution function, the following equations must hold for any x, t and τ :

$$(1) \quad \lim_{y \rightarrow -\infty} F(t, x; \tau, y) = 0, \quad \lim_{y \rightarrow +\infty} F(t, x; \tau, y) = 1;$$

(2) the function $F(t, x; \tau, y)$ is continuous from the left with respect to the argument y .

Now suppose that the function $F(t, x; \tau, y)$ is continuous with respect to t, τ and with respect to x .

Consider the instants of time t, s, τ ($t < s < \tau$). Since the system passes from state x at time t to one of the states of the interval $(z, z+dz)$ at time s with probability $d_z F(t, x; s, z)$ and from state z at time s to a state less than y at time τ with probability $F(s, z; \tau, y)$, we find, from formula (1) of Sec. 52, that

$$F(t, x; \tau, y) = \int F(s, z; \tau, y) d_z F(t, x; s, z)$$

It is natural to call this equality the *generalized Markov equation*, for it represents an extension of Equation (1), Sec. 17, of the theory of Markov chains to the theory of stochastic processes and in this theory plays just as important a part as the aforementioned identity does in the theory of Markov chains.

The probability $F(t, x; \tau, y)$ is so far defined only for $\tau > t$. We extend this definition by taking

$$\begin{aligned} \lim_{\tau \rightarrow t+0} F(t, x; \tau, y) &= \lim_{t \rightarrow \tau-0} F(t, x; \tau, y) = E(x, y) = \\ &= \begin{cases} 0 & \text{for } y \leq x \\ 1 & \text{for } y > x \end{cases} \end{aligned}$$

* Note that the parameter t (time) is ordinarily specified on the half-line ($t \geq t_0$).

If there exists a density

$$f(t, x; \tau, y) = \frac{\partial}{\partial y} F(t, x; \tau, y)$$

the obvious equalities

$$\int_{-\infty}^y f(t, x; \tau, x) dz = F(t, x; \tau, y)$$

$$\int f(t, x; \tau, z) dz = 1$$

hold.

For this case, the generalized Markov equation should be written in the form

$$f(t, x; \tau, y) = \int f(s, z; \tau, y) f(t, x; s, z) dz$$

Sec. 54. Continuous Stochastic Processes. Kolmogorov's Equations

We say that a stochastic process $\xi(t)$ is *continuous* if during small time intervals appreciable increments are obtainable only with a small probability $\xi(t)$. Here we demand that the process $\xi(t)$ be more strongly continuous, namely: no matter what the constant δ ($\delta > 0$), the following relation holds:

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|y-x| \geq \delta} d_y F(t - \Delta t, x; t, y) = 0 \quad (1)$$

Our immediate task is to derive the differential equations which, upon fulfillment of certain conditions, will be satisfied by the function $F(t, x; \tau, y)$ that governs a continuous stochastic process without aftereffect. These equations were first proved in rigorous fashion by A. N. Kolmogorov (though the second one had occurred prior in the works of physicists) and are called *Kolmogorov's equations*.

We assume that

(1) the partial derivatives

$$\frac{\partial F(t, x; \tau, y)}{\partial x} \quad \text{and} \quad \frac{\partial^2 F(t, x; \tau, y)}{\partial x^2}$$

exist and are continuous for all values of t, x, y and $\tau > t$;

(2) for arbitrary $\delta > 0$ there exist the limits

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|y-x| < \delta} (y-x) d_y F(t - \Delta t, x; t, y) = a(t, x) \quad (2)$$

and

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|y-x| < \delta} (y-x)^2 d_y F(t-\Delta t, x; t, y) = b(t, x) \quad (3)$$

and this convergence is uniform in x . *

The left-hand sides of (2) and (3) depend on δ . However, this dependence, by virtue of the definition of the continuity of a process (that is, by virtue of (1)), is only apparent.

Kolmogorov's First Equation. *If the foregoing conditions (1) and (2) are fulfilled, the function $F(t, x; \tau, y)$ satisfies the equation*

$$\frac{\partial F(t, x; \tau, y)}{\partial t} = -a(t, x) \frac{\partial F(t, x; \tau, y)}{\partial x} - \frac{b(t, x)}{2} \frac{\partial^2 F(t, x; \tau, y)}{\partial x^2} \quad (4)$$

Proof. According to the generalized Markov equation,

$$F(t-\Delta t, x; \tau, y) = \int F(t, z; \tau, y) d_z F(t-\Delta t, x; t, z)$$

Also, by virtue of the properties of a distribution function,

$$F(t, x; \tau, y) = \int F(t, x; \tau, y) d_z F(t-\Delta t, x; t, z)$$

From these equations we conclude that

$$\begin{aligned} \frac{F(t-\Delta t, x; \tau, y) - F(t, x; \tau, y)}{\Delta t} &= \\ &= \frac{1}{\Delta t} \int [F(t, z; \tau, y) - F(t, x; \tau, y)] d_z F(t-\Delta t, x; t, z) \end{aligned}$$

* A. N. Kolmogorov proved the existence of the limits $a(t, x)$ and $b(t, x)$ on the assumption that for given x and s the determinant

$$\begin{vmatrix} \frac{\partial}{\partial x} f(s, x, t', y') & \frac{\partial}{\partial x} f(s, x, t'', y'') \\ \frac{\partial^2}{\partial x^2} f(s, x, t', y') & \frac{\partial^2}{\partial x^2} f(s, x, t'', y'') \end{vmatrix}$$

does not vanish identically for arbitrary t', t'', y', y'' . Literally repeating Kolmogorov's reasoning, one can prove that from (1) and upon the assumption that for given x and s the determinant

$$\begin{vmatrix} \frac{\partial}{\partial x} F(s, x, t', y') & \frac{\partial}{\partial x} F(s, x, t'', y'') \\ \frac{\partial^2}{\partial x^2} F(s, x, t', y') & \frac{\partial^2}{\partial x^2} F(s, x, t'', y'') \end{vmatrix}$$

does not vanish identically for arbitrary t', t'', y', y'' , there follows the existence of the limits $a(t, x)$ and $b(t, x)$.

At the end of this section we will find out what physical meaning the functions a and b have.

By Taylor's formula, given the assumptions we have made, the following equality holds:

$$F(t, z; \tau, y) = F(t, x; \tau, y) + (z-x) \frac{\partial F(t, x; \tau, y)}{\partial x} + \\ + \frac{1}{2} (z-x)^2 \frac{\partial^2 F(t, x; \tau, y)}{\partial x^2} + o((z-x)^2)$$

The following analytical transformations do not require any explanations:

$$\frac{F(t-\Delta t, x; \tau, y) - F(t, x; \tau, y)}{\Delta t} = \\ = \frac{1}{\Delta t} \int_{|z-x| \geq \delta} [F(t, z; \tau, y) - F(t, x; \tau, y)] d_z F(t-\Delta t, x; t, z) + \\ + \frac{1}{\Delta t} \int_{|z-x| < \delta} [F(t, z; \tau, y) - F(t, x; \tau, y)] d_z F(t-\Delta t, x; t, z) = \\ = \frac{1}{\Delta t} \int_{|z-x| \geq \delta} [F(t, z; \tau, y) - F(t, x; \tau, y)] d_z F(t-\Delta t, x; t, z) + \\ + \frac{\partial F(t, x; \tau, y)}{\partial x} \cdot \frac{1}{\Delta t} \int_{|z-x| < \delta} (z-x) d_z F(t-\Delta t, x; t, z) + \\ + \frac{1}{2} \frac{\partial^2 F(t, x; \tau, y)}{\partial x^2} \times \\ \times \frac{1}{\Delta t} \int_{|z-x| < \delta} [(z-x)^2 + o(z-x)^2] d_z F(t-\Delta t, x; t, z) \quad (5)$$

We now pass to the limit, letting $\Delta t \rightarrow 0$. The first term on the right-hand side, by virtue of (1), has the limit 0. The second term, by (2), has the limit $a(t, x) \frac{\partial F}{\partial x}$. Finally, the third term can differ from $\frac{1}{2} b(t, x) \frac{\partial^2 F}{\partial x^2}$ only by a summand that tends to zero as $\delta \rightarrow 0$. But since the left-hand side of the equality is independent of δ and the limiting values just mentioned are independent of δ , the limit of the right-hand side exists and is equal to

$$a(t, x) \frac{\partial F(t, x; \tau, y)}{\partial x} + \frac{1}{2} b(t, x) \frac{\partial^2 F(t, x; \tau, y)}{\partial x^2}$$

From this we conclude that the limit

$$\lim_{\Delta t \rightarrow 0} \frac{F(t-\Delta t, x; \tau, y) - F(t, x; \tau, y)}{\Delta t} = - \frac{\partial F(t, x; \tau, y)}{\partial t}$$

exists.

Equality (5) leads us to Equation (4).

If it is assumed that a density function

$$f(t, x; \tau, y) = \frac{\partial}{\partial y} F(t, x; \tau, y)$$

exists, then simple differentiation of (4) shows that the density $f(t, x; \tau, y)$ satisfies the equation

$$\frac{\partial f(t, x; \tau, y)}{\partial t} + a(t, x) \frac{\partial f(t, x; \tau, y)}{\partial x} + \frac{1}{2} b(t, x) \frac{\partial^2 f(t, x; \tau, y)}{\partial x^2} = 0 \quad (4')$$

Let us now derive Kolmogorov's second equation. In doing so we will not strive for the greatest possible generality and will make assumptions that are not required by the essence of the matter. Besides the assumptions that have already been made we impose on the function $F(t, x; \tau, y)$ the following additional restrictions:

(3) there exists a probability density function

$$f(t, x; \tau, y) = \frac{\partial F(t, x; \tau, y)}{\partial y}$$

(4) there exist the continuous derivatives

$$\frac{\partial f(t, x; \tau, y)}{\partial \tau}, \quad \frac{\partial}{\partial y} [a(\tau, y) f(t, x; \tau, y)], \quad \frac{\partial^2}{\partial y^2} [b(\tau, y) f(t, x; \tau, y)]$$

Kolmogorov's Second Equation*. *If conditions 1 to 4 are fulfilled, then for a continuous stochastic process without aftereffect the density $f(t, x; \tau, y)$ satisfies the equation*

$$\begin{aligned} \frac{\partial f(t, x; \tau, y)}{\partial \tau} = & - \frac{\partial}{\partial y} [a(\tau, y) f(t, x; \tau, y)] + \\ & + \frac{1}{2} \frac{\partial^2}{\partial y^2} [b(\tau, y) f(t, x; \tau, y)] \end{aligned} \quad (6)$$

Proof. Let a and b ($a < b$) be some numbers and $R(y)$ be a non-negative continuous function with continuous derivatives up to second order inclusive. Besides, we will demand that

$$R(y) = 0 \text{ for } y < a \text{ and } y > b$$

From the condition of continuity of the function $R(y)$ and its derivatives we conclude that

$$R(a) = R(b) = R'(a) = R'(b) = R''(a) = R''(b) = 0 \quad (7)$$

* Kolmogorov's second equation was earlier derived by the physicists Fokker and Planck in connection with the development of diffusion theory.

First, note that

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy &= \frac{\partial}{\partial \tau} \int_a^b f(t, x; \tau, y) R(y) dy = \\ &= \lim_{\Delta \tau \rightarrow 0} \int \frac{f(t, x; \tau + \Delta \tau, y) - f(t, x; \tau, y)}{\Delta \tau} R(y) dy \end{aligned}$$

According to the generalized Markov equation,

$$f(t, x; \tau + \Delta \tau, y) = \int f(t, x; \tau, z) f(\tau, z; \tau + \Delta \tau, y) dz$$

and so

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy &= \\ &= \lim_{\Delta \tau \rightarrow 0} \frac{1}{\Delta \tau} \left[\int \int f(t, x; \tau, z) f(\tau, z; \tau + \Delta \tau, y) R(y) dz dy - \right. \\ &\quad \left. - \int f(t, x; \tau, y) R(y) dy \right] = \\ &= \lim_{\Delta \tau \rightarrow 0} \frac{1}{\Delta \tau} \left[\int f(t, x; \tau, z) \int f(\tau, z; \tau + \Delta \tau, y) R(y) dy dz - \right. \\ &\quad \left. - \int f(t, x; \tau, y) R(y) dy \right] = \\ &= \lim_{\Delta \tau \rightarrow 0} \frac{1}{\Delta \tau} \int f(t, x; \tau, y) \left[\int f(\tau, y; \tau + \Delta \tau, z) R(z) dz - R(y) \right] dy \end{aligned}$$

The transformations that have been performed are obvious: the first time we changed the order of integration and the second time we changed the notation of the variables of integration (we replaced y by z and z by y).

By Taylor's formula

$$R(z) = R(y) + (z - y) R'(y) + \frac{1}{2} (z - y)^2 R''(y) + o[(z - y)^2]$$

Since, by virtue of the boundedness of the function $R(z)$ and the condition (1),

$$\int_{|y-z| \geq \delta} f(\tau, y; \tau + \Delta \tau, z) R(z) dz = o(\Delta \tau)$$

and

$$\int_{|y-z| < \delta} f(\tau, y; \tau + \Delta \tau, z) dz = 1 + o(\Delta \tau)$$

it follows that

$$\begin{aligned} \int f(\tau, y; \tau + \Delta\tau, z) R(z) dz - R(y) &= \\ &= R'(y) \int_{|y-z| < \delta} (z-y) f(\tau, y; \tau + \Delta\tau, z) dz + \\ &+ \frac{1}{2} R''(y) \int_{|y-z| < \delta} [(z-y)^2 + o(z-y)^2] f(\tau, y; \tau + \Delta\tau, z) dz + o(\Delta\tau) \end{aligned}$$

Thus,

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy &= \\ &= \lim_{\Delta\tau \rightarrow 0} \int_a^b f(t, x; \tau, y) \left\{ R'(y) \int_{|y-z| < \delta} (z-y) f(\tau, y; \tau + \Delta\tau, z) dz + \right. \\ &+ \frac{1}{2} R''(y) \int_{|y-z| < \delta} [(z-y)^2 + o(z-y)^2] f(\tau, y; \tau + \Delta\tau, z) dz + \\ &\quad \left. + o(\Delta\tau) \right\} dy \end{aligned}$$

We pass to the limit, letting $\Delta\tau \rightarrow 0$. By the assumption of uniform convergence to the limits in (2) and (3) we conclude that the preceding equality may be written in the form

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy &= \\ &= \int_a^b f(t, x; \tau, y) \left[a(\tau, y) R'(y) + \frac{1}{2} b(\tau, y) R''(y) \right] dy \end{aligned}$$

Since $R'(y) = R''(y) = 0$ for $y \leq a$ and $y \geq b$, it follows that

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy &= \\ &= \int_a^b f(t, x; \tau, y) \left[a(\tau, y) R'(y) + \frac{1}{2} b(\tau, y) R''(y) \right] dy \quad (8) \end{aligned}$$

Taking advantage of the formula of integration by parts and of equalities (7), we find

$$\begin{aligned} \int_a^b f(t, x; \tau, y) a(\tau, y) R'(y) dy &= - \int_a^b R(y) \frac{\partial}{\partial y} [a(\tau, y) f(t, x; \tau, y)] dy \\ \int_a^b f(t, x; \tau, y) b(\tau, y) R''(y) dy &= \int_a^b R(y) \frac{\partial^2}{\partial y^2} [b(\tau, y) f(t, x; \tau, y)] dy \end{aligned}$$

Substituting the expressions thus obtained into (8), we get

$$\begin{aligned} \int_a^b \frac{\partial f(t, x; \tau, y)}{\partial \tau} R(y) dy = \\ = \int_a^b \left\{ -\frac{\partial}{\partial y} [a(\tau, y) f(t, x; \tau, y)] + \right. \\ \left. + \frac{1}{2} \frac{\partial^2}{\partial y^2} [b(\tau, y) f(t, x; \tau, y)] \right\} R(y) dy \end{aligned}$$

This equation can obviously be written as

$$\begin{aligned} \int_a^b \left\{ \frac{\partial f(t, x; \tau, y)}{\partial \tau} + \frac{\partial}{\partial y} [a(\tau, y) f(t, x; \tau, y)] - \right. \\ \left. - \frac{1}{2} \frac{\partial^2}{\partial y^2} [b(\tau, y) f(t, x; \tau, y)] \right\} R(y) dy = 0 \quad (9) \end{aligned}$$

Since the function $R(y)$ is arbitrary, (6) follows from the last identity. Indeed, suppose that this is not so. Then there exists a number quadruple $(t, x; \tau, y)$ such that the expression in the braces of (9) is different from zero. By virtue of the assumptions made this expression is a continuous function; hence, there will be an interval $\alpha < y < \beta$ where the expression retains its sign. If $a \leq \alpha$ and $b \geq \beta$, then we assume $R(y) = 0$ for $y \leq \alpha$ and $y \geq \beta$ and $R(y) > 0$ for $\alpha < y < \beta$. Given this choice of $R(y)$, the integral on the left-hand side of (9) must be different from zero. We arrive at a contradiction. Thus, our assumption is erroneous and, hence, (6) follows from (9).

Naturally, the basic problem that has to be solved does not consist in verifying that the given function $f(t, x; \tau, y)$ satisfies the Kolmogorov equations, but in seeking an unknown function $f(t, x; \tau, y)$ on the basis of these equations in which the coefficients $a(t, x)$ and $b(t, x)$ are assumed to be known. What is sought here is not, of course, just any solution of the Kolmogorov equations, but only those which satisfy the following requirements:

$$\left. \begin{aligned} 1. & f(t, x; \tau, y) \geq 0 \text{ for all } t, x, \tau, y, \\ 2. & \int f(t, x; \tau, y) dy = 1 \\ \text{and for any } \delta > 0 \\ 3. & \lim_{\tau \rightarrow t} \int_{|y-x| \geq \delta} f(t, x; \tau, y) dy = 0 \end{aligned} \right\} \quad (10)$$

We shall not undertake to clarify the conditions that must be imposed on the functions $a(t, x)$ and $b(t, x)$ for a solution of the Kolmogorov equations to exist that would satisfy the enumerated requirements and would also be unique.

We shall slightly strengthen the requirement of continuity in order to elucidate the physical meaning of the coefficients $a(t, x)$ and $b(t, x)$: in place of (1) we assume that for any $\delta > 0$ the following relation holds:

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int_{|y-x| > \delta} (y-x)^2 d_y F(t-\Delta t, x; t, y) = 0 \quad (1')$$

It is easy to see that (1) follows from (1'). The requirements 2 and 3 can now be written differently, namely:

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int (y-x) d_y F(t-\Delta t, x; t, y) = a(t, x) \quad (2')$$

and

$$\lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} \int (y-x)^2 d_y F(t-\Delta t, x; t, y) = b(t, x) \quad (3')$$

The other requirements and also the final conclusions due to substitution of (1') for (1) remain unchanged. Since

$$\int (y-x) d_y F(t-\Delta t, x; t, y) = \mathbf{M} [\xi(t) - \xi(t-\Delta t)]$$

is the expectation of variation of $\xi(t)$ during time Δt , and

$$\int (y-x)^2 d_y F(t-\Delta t, x; t, y) = \mathbf{M} [\xi(t) - \xi(t-\Delta t)]^2$$

is the expectation of the square of the variation of $\xi(t)$ and, consequently, is proportional to the kinetic energy (under the assumption that $\xi(t)$ is the coordinate of a point moving under the effect of chance actions), it is clear from (2') and (3') that $a(t, x)$ is the average rate of change of $\xi(t)$, and $b(t, x)$ is proportional to the mean kinetic energy of the system under study.

We conclude this section with a consideration of a special case of the Kolmogorov equations when the function $f(t, x; \tau, y)$ depends on t, τ and $y-x$, but not on x and y themselves. Physically, this means that the process is homogeneous in space: the probability of the increment $\Delta = y-x$ is independent of the position x that the system was in at time t . Obviously, in this case the functions $a(t, x)$ and $b(t, x)$ do not depend on x and are functions solely of the argument t :

$$a(t) = a(t, x), \quad b(t) = b(t, x)$$

In our case, the Kolmogorov equations may be rewritten as

$$\left. \begin{aligned} \frac{\partial f}{\partial t} &= -a(t) \frac{\partial f}{\partial x} - \frac{1}{2} b(t) \frac{\partial^2 f}{\partial x^2} \\ \frac{\partial f}{\partial \tau} &= -a(\tau) \frac{\partial f}{\partial y} + \frac{1}{2} b(\tau) \frac{\partial^2 f}{\partial y^2} \end{aligned} \right\} \quad (11)$$

We first consider the special case when $a(t) = 0$ and $b(t) = 1$. Then Equations (11) become the equation of heat conduction

$$\left. \begin{aligned} \frac{\partial f}{\partial \tau} &= \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \\ \frac{\partial f}{\partial t} &= -\frac{1}{2} \frac{\partial^2 f}{\partial x^2} \end{aligned} \right\} \quad \text{and its adjoint} \quad (12)$$

From the general theory of the heat-conduction equation it is known that the only solution of these equations that satisfies the conditions (10) is given by the function

$$f(t, x; \tau, y) = \frac{1}{\sqrt{2\pi(\tau-t)}} e^{-\frac{(y-x)^2}{2(\tau-t)}}$$

The change of variables

$$\begin{aligned} x' &= x - \int_a^t a(z) dz, & y' &= y - \int_a^\tau b(z) dz \\ t' &= \int_a^t b(z) dz, & \tau' &= \int_a^\tau b(z) dz \end{aligned}$$

reduces (11) to Equations (12). This makes it possible to write the desired solution of Equations (11) as

$$f(t, x; \tau, y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-x-A)^2}{2\sigma^2}}$$

where

$$A = \int_t^\tau a(z) dz, \quad \sigma^2 = \int_t^\tau b(z) dz$$

Sec. 55. Purely Discontinuous Stochastic Processes. The Kolmogorov-Feller Equations

In modern natural science an important role is played by processes in which changes occur in a system in jumps and not continuously. Some problems of this nature were given in Sec. 50.

We say that a stochastic process $\xi(t)$ is *purely discontinuous* if in the course of some time interval $(t, t + \Delta t)$ the quantity $\xi(t)$ remains unchanged and equal to x with probability $1 - p(t, x) \Delta t + o(\Delta t)$ and can undergo change only with probability $p(t, x) \Delta t + o(\Delta t)$ [here we assume that the probability of more than one change of $\xi(t)$ during the time interval Δt is $o(\Delta t)$]. Quite naturally, since we confine ourselves to processes without aftereffect, the distribution function of

changes of $\xi(t)$ following a jump is no longer dependent on the value that $\xi(t)$ had at various times prior to the jump.

Denote by $P(t, x, y)$ the conditional distribution function of $\xi(t)$ provided that a jump occurred at time t and that immediately prior to the jump, $\xi(t)$ was equal to x (that is, $\xi(t-0)=x$).

The distribution function $F(t, x; \tau, y)$ can readily be expressed in terms of the function $p(t, x)$ and $P(t, x, y)$, namely:

$$F(t, x; \tau, y) = [1 - p(t, x)(\tau - t)] E(x, y) + (\tau - t) p(t, x) P(t, x, y) + o(\tau - t) \quad (1)$$

According to definition, the functions $p(t, x)$ and $P(t, x, y)$ are nonnegative; and since $P(t, x, y)$ is a distribution function, the equalities

$$P(t, x, -\infty) = 0, \quad P(t, x, +\infty) = 1$$

hold.

Besides, we will assume that $p(t, x)$ is bounded and that both the functions $p(t, x)$ and $P(t, x, y)$ are continuous in t and x (actually, it is sufficient to assume that they are Borel measurable in x).

We make no assumptions with respect to the function $F(t, x; \tau, y)$ and only retain its definition for $t = \tau$:

$$\begin{aligned} \lim_{\tau \rightarrow t+0} F(t, x; \tau, y) &= \lim_{t \rightarrow \tau-0} F(t, x; \tau, y) = E(x, y) = \\ &= \begin{cases} 0 & \text{for } y \leq x \\ 1 & \text{for } y > x \end{cases} \end{aligned}$$

One of the problems of this section is to prove the following theorem.

Theorem. *The distribution function $F(t, x; \tau, y)$ of a purely discontinuous process without aftereffect satisfies the following two integro-differential equations:*

$$\begin{aligned} \frac{\partial F(t, x; \tau, y)}{\partial t} &= p(t, x) [F(t, x; \tau, y) - \\ &\quad - \int F(t, z; \tau, y) d_z P(t, x, z)] \quad (2) \end{aligned}$$

$$\begin{aligned} \frac{\partial F(t, x; \tau, y)}{\partial \tau} &= - \int_{-\infty}^y p(t, z) d_z F(t, x; \tau, y) + \\ &\quad + \int p(\tau, z) P(\tau, z, y) d_z F(t, x; \tau, z) \quad (3) \end{aligned}$$

Equation (2) was obtained by A. N. Kolmogorov in 1931; under our assumptions, both Equations (2) and (3) were obtained by W. Feller in 1937. It is therefore natural to call Equations (2) and (3) *Kolmogorov-Feller equations*.

Proof. By virtue of the generalized Markov equation,

$$F(t, x; \tau, y) = \int F(t + \Delta t, z; \tau, y) d_z F(t, x; t + \Delta t, z)$$

Substituting the value of $F(t, x; t + \Delta t, z)$ from formula (1), we find

$$\begin{aligned} F(t, x; \tau, y) &= \\ &= \int F(t + \Delta t, z; \tau, y) d_z [1 - p(t, x) \Delta t + o(\Delta t)] E(x, z) + \\ &\quad + \int F(t + \Delta t, z; \tau, y) d_z [p(t, x) \Delta t + o(\Delta t)] P(t, x, z) \end{aligned}$$

Since

$$\int F(t + \Delta t, z; \tau, y) d_z E(x, z) = F(t + \Delta t, x; \tau, y)$$

it follows that

$$\begin{aligned} F(t, x; \tau, y) &= [1 - p(t, x) \Delta t] F(t + \Delta t, x; \tau, y) + \\ &\quad + \Delta t p(t, x) \int F(t + \Delta t, z; \tau, y) d_z P(t, x, z) + o(\Delta t) \end{aligned}$$

And from this we get

$$\begin{aligned} \frac{F(t + \Delta t, x; \tau, y) - F(t, x; \tau, y)}{\Delta t} &= p(t, x) F(t + \Delta t, x; \tau, y) - \\ &\quad - p(t, x) \int F(t + \Delta t, z; \tau, y) d_z P(t, x, z) + o(1) \end{aligned}$$

Passing to the limit gives us (2).

The Markov equation and (1), and also the definition of the function $E(x, z)$ enable us to write the following chain of equalities:

$$\begin{aligned} F(t, x; \tau + \Delta \tau, y) &= \int F(\tau, z; \tau + \Delta \tau, y) d_z F(t, x; \tau, z) = \\ &= \int \{ [1 - p(\tau, z) \Delta \tau] E(z, y) + \Delta \tau p(\tau, z) P(\tau, z, y) + o(\Delta \tau) \} \times \\ &\quad \times d_z F(t, x; \tau, z) = \\ &= \int_{-\infty}^y d_z F(t, x; \tau, z) - \Delta \tau \int_{-\infty}^y p(\tau, z) d_z F(t, x; \tau, z) + \\ &\quad + \Delta \tau \int p(\tau, z) P(\tau, z, y) d_z F(t, x; \tau, z) + o(\Delta \tau) \end{aligned}$$

Equation (3) and the existence of the derivative $\frac{\partial F}{\partial \tau}$ follow from this in the ordinary way.

We shall solve yet another important problem in applications: What is the probability of a system changing its state n times ($n = 0, 1, 2, \dots$) during a time interval from t to τ ($\tau > t$)?

Denote by $p_n(t, x, \tau)$ the probability that a system starting from state x at time t will change its state n times up to time τ . We begin the solution with $n=0$.

Write the following equation:

$$p_0(t, x, \tau) = p_0(t, x, \tau + \Delta\tau) + p_0(t, x, \tau) [1 - p_0(\tau, x, \tau + \Delta\tau)] \quad (4)$$

which states that the absence of changes in the state of the system during the time interval (t, τ) can take place in two mutually exclusive ways: (1) the system has not changed its state during a large interval of time $(t, \tau + \Delta\tau)$; (2) the system did not change its state prior to time τ , but during the time interval $(\tau, \tau + \Delta\tau)$ its state changed. Since by definition of a purely discontinuous process

$$p_0(\tau, x, \tau + \Delta\tau) = 1 - p(\tau, x) \Delta\tau + o(\Delta\tau)$$

equation (4) may be written otherwise:

$$\frac{p_0(t, x, \tau + \Delta\tau) - p_0(t, x, \tau)}{\Delta\tau} = -p_0(t, x, \tau) p(\tau, x) + o(1)$$

Whence, letting $\Delta\tau \rightarrow 0$, we find that the derivative $\frac{\partial p_0(t, x, \tau)}{\partial \tau}$ exists and that

$$\frac{\partial p_0(t, x, \tau)}{\partial \tau} = -p_0(t, x, \tau) p(\tau, x)$$

Integrating this equation, we find

$$p_0(t, x, \tau) = C e^{-\int_t^\tau p(u, x) du}$$

Since

$$p_0(\tau, x, \tau) = 1$$

it follows that $C=1$ and

$$p_0(t, x, \tau) = e^{-\int_t^\tau p(u, x) du} \quad (5)$$

We will now see that knowing $p_0(t, x, \tau)$ and also the function $P(t, x, y)$ defined earlier, we can calculate any probability $p_n(t, x, \tau)$. Indeed, an n -fold change of the state occurs in the following manner:

(1) prior to time s ($t < s < \tau$) the system does not change its state [the probability of this event is equal to $p_0(t, x, s)$];

(2) during the interval $(s, s + \Delta s)$ the system changes its state [the probability of this is equal to $p_1(s, x, s + \Delta s) = p(s, x) \Delta s + o(\Delta s)$];

(3) the probability that the new state at which the system will arrive will lie between y and $y + \Delta y$ is equal to

$$P(s, x, y + \Delta y) - P(s, x, y) = \Delta_y P(s, x, y)$$

(4) finally, during time $(s+\Delta s, \tau)$ the system will change its state $n-1$ times [the probability of this event is $p_{n-1}(s+\Delta s, y, \tau)$].

The probability that all four of these events will occur is, by the multiplication theorem, equal to

$$p_0(t, x, s) [p(s, x) + o(1)] \Delta s \cdot \Delta_y P(s, x, y) \cdot p_{n-1}(s+\Delta s, y, \tau)$$

Since s and y may be arbitrary ($t < s < \tau$ and $-\infty < y < \infty$), then by the formula of total probability

$$\begin{aligned} p_n(t, x, \tau) &= \int_t^\tau \int p_0(t, x, s) p(s, x) p_{n-1}(s, y, \tau) d_y P(s, x, y) ds = \\ &= \int_t^\tau p_0(t, x, s) p(s, x) \int p_{n-1}(s, y, \tau) d_y P(s, x, y) ds \quad (6) \end{aligned}$$

Whence, in particular,

$$p_1(t, x, \tau) = \int_t^\tau p_0(t, x, s) p(s, x) \int p_0(s, y, \tau) d_y P(s, x, y) ds \quad (7)$$

The procedure for determining $p_n(t, x, \tau)$ is obvious; by formula (5) we find $p_0(t, x, \tau)$, by formula (7) we compute $p_1(t, x, \tau)$ and in succession, $p_2(t, x, \tau)$, $p_3(t, x, \tau)$ and, finally, $p_n(t, x, \tau)$.

Example 1. Let the variable $\xi(t)$ that interests us be the number of changes of the state during time from 0 to τ . Assuming that $p(t, x) = a$, where a is a positive constant, find $p_n(t, x, \tau)$.

In our case, possible states of the system will be all the nonnegative integers ($x=0, 1, 2, \dots$) and only these integers. Since in each change of state the variable $\xi(t)$ increases exactly by 1, it follows that

$$P(t, x, y) = \begin{cases} 0 & \text{for } y \leq x \\ 1 & \text{for } y > x \end{cases}$$

From formula (5) we have

$$p_0(t, x, \tau) = e^{-a(\tau-t)}$$

According to (7),

$$\begin{aligned} p_1(t, x, \tau) &= \int_t^\tau p_0(t, x, s) p(s, x) p_0(s, x+1, \tau) ds = \\ &= a \int_t^\tau e^{-(s-t)a} e^{-(\tau-s)a} ds = a(\tau-t) e^{-a(\tau-t)} \end{aligned}$$

By formula (6)

$$p_2(t, x, \tau) = \int_t^\tau p_0(t, x, s) p(s, x) p_1(s, x+1, \tau) ds = \frac{[a(\tau-t)]^2}{2!} e^{-a(\tau-t)}$$

Now suppose that

$$p_{n-1}(t, x, \tau) = \frac{[a(\tau-t)]^{n-1}}{(n-1)!} e^{-a(\tau-t)}$$

By formula (6)

$$\begin{aligned} p_n(t, x, \tau) &= \int_t^\tau p_0(t, x, s) p(s, x) p_{n-1}(s, x+1, \tau) ds = \\ &= t \int_t^\tau \frac{a [a(\tau-s)]^{n-1}}{(n-1)!} e^{-a(\tau-s)} ds = \frac{[a(\tau-t)]^n}{n!} e^{-a(\tau-t)} \end{aligned}$$

This proves that for any integer $n \geq 0$,

$$p_n(t, x, \tau) = \frac{[a(\tau-t)]^n}{n!} e^{-a(\tau-t)}$$

The solution of our problem is therefore the Poisson law. In particular,

$$p_n(0, 0, \tau) = \frac{(a\tau)^n}{n!} e^{-a\tau}$$

It is easy to see that the function

$$F(t, x; \tau, y) = \begin{cases} 0 & \text{for } y \leq 0 \\ \sum_{n < y} \frac{[a(\tau-t)]^n}{n!} e^{-a(\tau-t)} & \text{for } y > 0 \end{cases}$$

is the solution of integro-differential Equations (2) and (3).

Example 2. At time $t=0$ there are N radioactive atoms. The probability of decay of an atom in the time interval $(t, t+\Delta t)$ is equal to $aN(t)\Delta t + o(\Delta t)$, where $a > 0$ is a constant and $N(t)$ is the number of atoms that have not decayed up to time t . Find the probability that during the time between t and τ there will be n disintegrations*.

This is a typical purely discontinuous stochastic process. The quantity n can, quite understandably, take only the values $0, 1, 2, \dots, N(t)$.

* We assume here that the decay products do not disintegrate and at any rate do not affect the atoms that have not yet disintegrated.

By the condition of the problem

$$p(t, x) = \begin{cases} 0 & \text{for } x \leq 0 \text{ and } x \geq N \\ a(N-x) & \text{for } 0 < x \leq N \end{cases}$$

and

$$P(t, x, y) = \begin{cases} 0 & \text{for } y \leq x \\ 1 & \text{for } y > x \end{cases}$$

Let us first evaluate the probability that during the time from 0 to t there will be n disintegrations. By formula (5)

$$p_0(0, 0, \tau) = e^{-\int_0^\tau p(t, 0) dt} = e^{-aN\tau}$$

In exactly the same way we have

$$p_0(t, k, \tau) = e^{-a(N-k)(\tau-t)}$$

Then, by formula (7),

$$\begin{aligned} p_1(0, 0, \tau) &= \int_0^\tau p_0(0, 0, s) p(s, 0) p_0(s, 1, \tau) ds = \\ &= \int_0^\tau e^{-aN s} a N e^{-a(N-1)(\tau-s)} ds = \\ &= N e^{-aN\tau} \int_0^\tau a e^{a(\tau-s)} ds = N e^{-aN\tau} [e^{a\tau} - 1] \end{aligned} \quad (8)$$

By formula (6), it is easy to find, in succession, $p_2(0, 0, \tau)$, $p_3(0, 0, \tau)$, and so on and to prove that

$$p_n(0, 0, \tau) = C_N^n e^{-aN\tau} [e^{a\tau} - 1]^n \quad (9)$$

This is left to the reader.

Obviously, when $0 \leq n \leq N-k$ we have the following equation:

$$p_n(t, k, \tau) = C_{N-k}^n e^{-a(N-k)(\tau-t)} [e^{a(\tau-t)} - 1]^n \quad (9')$$

We are now in a position to determine the probability we are interested in. We denote it by $p_n(t, \tau)$. By the formula of total probability and then by using (9) and (9'), we find

$$\begin{aligned} p_n(t, \tau) &= \sum_{k=0}^{N-n} p_k(0, 0, t) p_n(t, k, \tau) = \\ &= \sum_{k=0}^{N-n} C_N^k e^{-aNt} [e^{at} - 1]^k C_{N-k}^n e^{-a(N-k)(\tau-t)} [e^{a(\tau-t)} - 1]^n = \\ &= e^{-aN\tau} [e^{a(\tau-t)} - 1]^n \sum_{k=0}^{N-n} C_N^k C_{N-k}^n e^{ak(\tau-t)} [e^{at} - 1]^k \end{aligned}$$

Since

$$C_N^k C_{N-k}^n = C_N^n C_{N-n}^k$$

and

$$\sum_{k=0}^{N-n} C_{N-n}^k [e^{a(\tau-t)} (e^{at} - 1)]^k = [1 + e^{a\tau} - e^{a(\tau-t)}]^{N-n}$$

it follows, finally, that

$$p_n(t, \tau) = C_N^n [e^{-at} - e^{-a\tau}]^n [e^{-a\tau} + 1 - e^{-at}]^{N-n}$$

It will readily be seen that the function

$$F(t, x; \tau, y) = \begin{cases} 0 & \text{for } y \leq x \\ \sum_{n < y} p_n(t, x, \tau) & \text{for } y < N - x \\ 1 & \text{for } y > N - x \end{cases}$$

is the solution of the Kolmogorov-Feller integro-differential equations.

Sec. 56. Homogeneous Stochastic Processes with Independent Increments

We now consider an important class of stochastic processes which will be fully described in terms of characteristic functions.

By a *homogeneous stochastic process with independent increments* is meant a collection of random variables $\xi(t)$ dependent on a single real parameter t and satisfying the following two conditions:

- (1) the distribution function of the variable $\xi(t+t_0) - \xi(t_0)$ is independent of t_0 (the process is time-homogeneous);
- (2) for any finite number of nonoverlapping intervals (a, b) of parameter t , the increments of the variable $\xi(t)$, that is, the differences $\xi(b) - \xi(a)$, are mutually independent (the increments are independent).

Of homogeneous stochastic processes with independent increments particular attention has been paid to the processes of *Brownian motion*, which later were also termed *Wiener processes*. For processes of this type, the following two conditions are assumed to be satisfied in addition to the two that have already been given:

- (3) the variables $\xi(t+t_0) - \xi(t_0)$ are normally distributed;

- (4) $M[\xi(t+t_0) - \xi(t_0)] = 0$

$D[\xi(t+t_0) - \xi(t_0)] = \sigma^2 t$ where σ^2 is a constant.

In Sec. 51 we considered another homogeneous process with independent increments, the Poisson process.

Before proceeding to obtain concrete results, let us consider several examples. In these examples, the foregoing conditions may be taken as a working hypothesis. Naturally, their sole justification is agreement of the conclusions with experiment.

Example 1. Diffusion of Gases. Consider a molecule of some gas that is moving among other molecules of the same gas under conditions of constant temperature and density. We introduce Cartesian coordinates and see how, in the course of time, one of the coordinates of the chosen molecule (say the x -coordinate) varies.

Because of random collisions of this molecule with the other molecules, the coordinate will vary in time as it receives random increments. The requirement that the conditions of the gas be constant obviously means that the process under study is homogeneous in time. In view of the large number of moving molecules and of the weak dependence of their motion, this is a process with independent increments.

Example 2. Molecular Speeds. Again consider a gas molecule moving in a volume of space filled with the molecules of some gas at constant density and temperature. Again refer the entire space to Cartesian coordinates and follow the time-variations of the velocity component along one of the coordinate axes. In its motion the molecule will be subject to random collisions with other molecules. The velocity component, due to these collisions, will receive random increments. Again we have a homogeneous stochastic process with independent increments.

Example 3. Radioactive Disintegration. The essence of radioactivity of a substance is known to consist in the fact that the atoms of a substance are converted into the atoms of another substance and considerable quantities of energy are released. Observations of comparatively large masses of a radioactive substance show that the various atoms disintegrate independently of one another so that the number of disintegrations of atoms in nonoverlapping intervals of time are mutually independent. Besides, the probabilities that during a time interval of definite length there will occur a certain number of disintegrations depend on the length of the interval and are practically independent of its location in time. Actually, of course, the radioactivity of the substance gradually falls off as its mass diminishes in size. However, for comparatively small time intervals (and for quantities of substance not excessively great) this change is so insignificant that it can definitely be neglected.

Any number of instances may be given in which the natural phenomenon or technological process that interests us is a homogeneous process with independent increments. A few examples are: cosmic radiation (the number of cosmic-ray particles impinging on a definite area during a specific time interval), the breaking of yarn on a ring spinning frame, the working load of a telephone operator (the number of calls in a certain interval of time) and so forth.

Let us now clarify the characteristic property of homogeneous stochastic processes with independent increments.

Denote by $F(x, \tau)$ the distribution function of the increment in the variable $\xi(t)$ during a time interval of duration τ . Then, if the time intervals of duration τ_1 and τ_2 do not overlap, we have

$$F(x; \tau_1 + \tau_2) = \int F(x - y; \tau_1) d_y F(y, \tau_2) \quad (1)$$

If $f(z, \tau)$ is the characteristic function, that is if

$$f(z, \tau) = \int e^{izx} d_x F(x; \tau)$$

it follows that Equation (1) will, in terms of characteristic functions, take on the following form:

$$f(z; \tau_1 + \tau_2) = f(z, \tau_1) f(z, \tau_2) \quad (1')$$

Generally speaking, if the time intervals $\tau_1, \tau_2, \dots, \tau_n$ do not overlap, then

$$f\left(z; \sum_{k=1}^n \tau_k\right) = \prod_{k=1}^n f(z; \tau_k)$$

In particular, if $\tau_1 = \tau_2 = \dots = \tau_n$ and $\sum_{k=1}^n \tau_k = \tau$ then

$$f(z; \tau) = \left[f\left(z; \frac{\tau}{n}\right) \right]^n$$

Thus, the *distribution function of any homogeneous stochastic process with independent increments is infinitely divisible.*

It should be pointed out that it was through the study of homogeneous processes with independent increments that infinitely divisible distribution laws began to be studied in the theory of probability. We have seen that the theory of infinitely divisible distribution laws exerted a decisive effect on the development of the classical problems of probability theory concerning the summation of random variables. As we have already pointed out, whereas before the interests of investigators centred on determining the broadest possible conditions for the law of large numbers and the convergence of normalized sums to the normal law, after A. N. Kolmogorov fully characterized the class of laws that govern homogeneous stochastic processes without aftereffect, it was quite natural for those general problems to emerge that were considered in the preceding chapter. And here it was found that the basic distribution laws, which earlier were obtained as asymptotic laws, in the theory of stochastic processes play the role of exact solutions of the appropriate functional equations. More than that, this new point of view permitted clarifying the causes by virtue of which in the classical theory of probability only two limiting distribution functions—the normal law and the Poisson law—were considered.

Since for arbitrary $\tau > 0$ in homogeneous processes with independent increments

$$f(z; \tau) = [f(z, 1)]^\tau$$

it follows that such processes are fully determined by specification of the characteristic function of the variable $\xi(1) - \xi(0)$. In Sec. 45 we saw that, for infinitely divisible laws with finite variance,

$$\log f(z; 1) = i\gamma z + \int \{e^{izu} - 1 - izu\} \frac{1}{u^2} dG(u) \quad (2)$$

where γ is a real constant and $G(u)$ is a nondecreasing function of bounded variation. We shall consider only this special case of homogeneous processes.

We introduce the following notations into formula (2):

$$\begin{aligned} M(u) &= \int_{-\infty}^u \frac{1}{x^2} dG(x) \quad \text{for } u < 0 \\ N(u) &= - \int_u^{\infty} \frac{1}{x^2} dG(x) \quad \text{for } u > 0 \\ \sigma^2 &= G(+0) - G(-0) \end{aligned}$$

Then it will take the form

$$\begin{aligned} \log f(z, 1) &= i\gamma z - \frac{\sigma^2 z^2}{2} + \int_{-\infty}^0 \{e^{izu} - 1 - izu\} dM(u) + \\ &\quad + \int_0^{\infty} \{e^{izu} - 1 - izu\} dN(u) \quad (2') \end{aligned}$$

Let us now clarify the probabilistic meaning of the functions $M(u)$ and $N(u)$.

In deriving the formulas of the canonical representation of infinitely divisible laws in Sec. 45, we introduced the function

$$G_n(u) = n \int_{-\infty}^u x^2 d\Phi_n(x)$$

We put

$$M_n(u) = \int_{-\infty}^u \frac{1}{x^2} dG_n(x) = n\Phi_n(u) \quad \text{for } u < 0$$

and

$$N_n(u) = - \int_u^{\infty} \frac{1}{x^2} dG_n(x) = -n[1 - \Phi_n(u)] \quad \text{for } u > 0$$

From the fact that as $n \rightarrow \infty$ at continuity points of the function $G(u)$

$$G_n(u) \rightarrow G(u)$$

we conclude from Helly's second theorem that at the continuity points of the function $M(u)$

$$M_n(u) = n\Phi_n(u) \rightarrow M(u)$$

From the viewpoint of stochastic processes, $\Phi_n(x)$ ($x < 0$) is the probability that the variable $\xi(\tau)$ will receive a negative increment greater than x in absolute value, in the interval $\left(\frac{k}{n}, \frac{k+1}{n}\right)$ of variation of the parameter τ . Thus, $M_n(x)$ is the sum over all k from 0 to $n-1$ of the probabilities that the variable $\xi(t)$ will acquire a negative increment (in jumps that are in absolute value greater than x) in the interval $\left(\frac{k}{n}, \frac{k+1}{n}\right)$ of variation of the parameter τ . Since $M(u)$ and $N(u)$ are the limits of the functions $M_n(u)$ and $N_n(u)$ respectively as $n \rightarrow \infty$, they are referred to as *jump functions*.

If $M(u) \equiv 0$ (for $u < 0$) and $N(u) \equiv 0$ (for $u > 0$), that is, there are no jump functions, then it will be seen from formula (2') that in this case the stochastic process is governed by the normal law. We see that a stochastic process governed by the normal law is continuous in the meaning of probability theory. We shall now prove a stronger assertion.

Theorem. *For a homogeneous stochastic process with independent increments and finite variance* to be governed by the normal law**, it is necessary and sufficient that for arbitrary $\varepsilon > 0$ the probability that the maximum of the absolute value of increments in $\xi(\tau)$ during the intervals $\left(\frac{k-1}{n}, \frac{k}{n}\right)$ ($k=1, 2, \dots, n$) will exceed ε should tend to zero together with $\frac{1}{n}$ ***.*

Proof. We have just seen that a homogeneous stochastic process with independent increments is governed by the normal law if and only if, for $x > 0$,

$$M(-x) \equiv N(x) \equiv 0 \quad (3)$$

Since

$$M(u) = \lim_{n \rightarrow \infty} M_n(u) \text{ and } N(u) = \lim_{n \rightarrow \infty} N_n(u)$$

* The theorem is true even without assuming the variance to be finite.

** In particular, by the normal law with variance zero, which is a law like $F(x) = 0$ for $x \leq a$, $F(x) = 1$ for $x > a$.

*** Thus, only processes governed by the normal law are "uniformly continuous" in the probabilistic sense.

it follows that the condition (3) is equivalent to the following:

$$\lim_{n \rightarrow \infty} n \Phi_n(-u) = \lim_{n \rightarrow \infty} n [1 - \Phi_n(u)] = 0 \quad (4)$$

We denote the increment $\xi(\tau)$ in the interval $\left(\frac{k-1}{n}, \frac{k}{n}\right)$ by ξ_{nk} ; then

$$p_{nk} = \Phi_n(-x) + 1 - \Phi_n(x+0) = \mathbf{P} \{ |\xi_{nk}| > x \}$$

It is obvious that the relations (4) are equivalent to the following:

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n p_{nk} = 0$$

From the inequalities

$$1 - \sum_{k=1}^n p_{nk} \leq \prod_{k=1}^n (1 - p_{nk}) \leq e^{-\sum_{k=1}^n p_{nk}} \leq 1$$

we see that the relations (4) are equivalent to the assertion that

$$\lim_{n \rightarrow \infty} \prod_{k=1}^n (1 - p_{nk}) = 1$$

which means that the probability that the inequalities $|\xi_{nk}| < \varepsilon$ will be realized for all k ($1 \leq k \leq n$), as $n \rightarrow \infty$, tends to unity. To put it otherwise, we have proved that the relations (3) hold if and only if, as $n \rightarrow \infty$,

$$\mathbf{P} \left\{ \max_{1 \leq k \leq n} |\xi_{nk}| \geq \varepsilon \right\} \rightarrow 0$$

which is what we set out to prove.

Sec. 57. The Concept of a Stationary Stochastic Process.

Khinchin's Theorem on the Correlation Coefficient

The Markovian processes, or processes without aftereffect, that we have studied in the preceding sections do not by any means exhaust the demands made by the natural sciences on probability theory. Indeed, in many cases earlier states of a system exert an extremely strong effect on the probability of its future states and this effect of the past cannot be dismissed even in an approximate interpretation of the problem. In principle, the situation can be corrected by changing the concept of the state of a system via the introduction of new parameters. For instance, if we considered a change in the position of a particle in diffusion processes or the Brownian motion as a process without aftereffect, this would mean that we disregard the inertia of the particle, which quite naturally plays an essential role in these

phenomena. In our example, the situation might be rectified by introducing the velocity of the particle into the concept of the state of the particle in addition to its coordinates. However, there are cases where this does not facilitate the solution of the problems. First of all, this refers to statistical mechanics, in which indication of the position of a point in one or another cell of phase space only yields a probabilistic judgement about its future state. Here, a knowledge of earlier positions of the point alters very essentially our judgements about the future of the point. In this connection, A. Ya. Khinchin isolated an important class of stochastic processes with aftereffect, the so-called *stationary processes*, which behave homogeneously in time.

A stochastic process $\xi(t)$ is called *stationary* if the n -dimensional distribution functions of the probabilities for two finite groups of variables $\xi(t_1), \xi(t_2), \dots, \xi(t_n)$ and $\xi(t_1+u), \xi(t_2+u), \dots, \xi(t_n+u)$ coincide and, hence, are independent of u . The numbers n and u , and also the instants of time t_1, t_2, \dots, t_n may be chosen here quite arbitrarily.

If we introduce the notation

$$F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) = \\ = P\{\xi(t_1) < x_1, \xi(t_2) < x_2, \dots, \xi(t_n) < x_n\}$$

then in accordance with the foregoing definition the following equation holds for any u and n :

$$F(x_1, x_2, \dots, x_n; t_1+u, t_2+u, \dots, t_n+u) = \\ = F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n) \quad (1)$$

The distribution functions $F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$ of any stochastic process must obviously satisfy the following two conditions:

(1) the *symmetry condition*: the equation

$$F(x_{i_1}, x_{i_2}, \dots, x_{i_n}; t_{i_1}, t_{i_2}, \dots, t_{i_n}) = \\ = F(x_1, x_2, \dots, x_n; t_1, t_2, \dots, t_n)$$

holds for any permutation i_1, i_2, \dots, i_n of the numbers $1, 2, \dots, n$;

(2) the *compatibility condition*: if $m < n$, then for any $t_{m+1}, t_{m+2}, \dots, t_n$

$$F(x_1, x_2, \dots, x_m; t_1, t_2, \dots, t_m) = \\ = F(x_1, x_2, \dots, x_m, \infty, \dots, \infty; t_1, t_2, \dots, t_m, t_{m+1}, \dots, t_n) \quad (2)$$

During recent years, the theory of stationary processes has found considerable applications in physics and engineering.

Stationary processes have been found to lie at the root of certain acoustic phenomena, including the random noises of radio engineering, and also of the search for hidden periodicities in astronomy, geophysics, and meteorology.

Steady-state technological processes frequently exhibit phenomena of the nature of stationary processes. As an illustration, consider the process of spinning. An appreciable inhomogeneity of the properties of spinning materials (length, strength and cross-section, etc., of the fibre), fluctuations in the rate and uniformity of feed of the product to the machines during various stages of the spinning process, and many other factors result in the properties of the yarn varying from one cross-sectional point to another. And it turns out that a knowledge of one or another property of the yarn in some part of the skein does not yield a complete knowledge of its properties in any other portion. But since the spinning process may be regarded as a steady-state process, the probabilistic characteristics of the quality of the yarn constitute a stationary process.

Clearly, any numerical characteristic of a stationary process $\xi(t)$ is independent of the time t and, for example, if $\xi(t)$ has a finite variance, then, obviously, the following equations hold:

$$\begin{aligned} M\xi(t+u) &= M\xi(t) = M\xi(0) = a \\ D\xi(t+u) &= D\xi(t) = D\xi(0) = \sigma^2 \\ M\{\xi(t+u)\xi(t)\} &= M\{\xi(u)\xi(0)\} \end{aligned}$$

This circumstance enables us, without restricting the generality of subsequent results, to assume $a=0$ and $\sigma=1$ [to achieve this it is obviously sufficient to consider the ratio $\frac{\xi(t)-a}{\sigma}$ in place of $\xi(t)$].

To take an important example, we consider a normal stationary process. For any n ($n=1, 2, \dots$) let the vector $\Xi_n = \{\xi(t_1), \xi(t_2), \dots, \xi(t_n)\}$ be normally distributed. We assume that

$$M\xi(t_j) = 0, \quad D\xi(t_j) = 1 \quad (-\infty < t_j < \infty)$$

and suppose that

$$M\xi(t_i)\xi(t_j) = R(t_i - t_j)$$

where $R(0)=1$ and $R(t)$ is an even function of t . The function $R(t)$ is such that the quadratic form

$$\sum_{i=1}^n \sum_{j=1}^n R(t_i - t_j) x_i x_j$$

is positive definite.

Since the characteristic function of the vector Ξ_n is

$$f_n(u_1, u_2, \dots, u_n; t_1, t_2, \dots, t_n) = \exp\left(-\sum_{i=1}^n \sum_{j=1}^n R(t_i - t_j) u_i u_j\right)$$

and the characteristic function of the vector $\Xi_k = \{\xi(t_1), \dots, \xi(t_k)\}$ is, for any $k < n$, equal to

$$f_k(u_1, \dots, u_k; t_1, \dots, t_k) = \exp\left(-\sum_{i=1}^k \sum_{j=1}^k R(t_i - t_j) u_i u_j\right) = \\ = f_n(u_1, \dots, u_k, 0, \dots, 0; t_1, \dots, t_k, t_{k+1}, \dots, t_n)$$

we conclude that the stochastic process we have defined satisfies a self-compatibility* condition. Besides, it is directly obvious that this process is stationary.

A homogeneous Markov process, that is, a Markov process for which the transition probability $F(t, x; \tau, y)$ is a function solely of the three arguments x , y and $\tau - t$, is also stationary.

In many problems of theory and in applications, multidimensional distributions of type (1) are not considered; and the only use made of the stationarity of the process is the constancy of the expectation, the variance, and the dependence of the correlation coefficient solely on the difference of the values of the parameter t . It is therefore natural to generalize the definition of stationarity and say that a *stochastic process is stationary in the broad sense* if the expectation and variance of $\xi(t)$ are independent of t , and the correlation coefficient of $\xi(t)$ and $\xi(t+u)$ is a function of u alone.

Clearly, a process $\xi(t)$ cannot be determined from a knowledge only of second moments and, consequently, such knowledge cannot completely take the place of the theory of stochastic processes based on a consideration of probability distributions. Nevertheless, in many problems the theory, which is based solely on a consideration of second moments (or, *correlation theory*, as it is called), proves sufficient and yields satisfactory results.

In this section we confine ourselves to a study of the *correlation function*, that is, the correlation coefficient of $\xi(t)$ and $\xi(t+u)$:

$$R(u) = \frac{\mathbf{M}[\xi(t+u) - \mathbf{M}\xi(t+u)] \mathbf{M}[\xi(t) - \mathbf{M}\xi(t)]}{\sqrt{\mathbf{D}\xi(t) \mathbf{D}\xi(t+u)}}$$

By virtue of our assumption that $a=0$ and $\sigma=1$, the expression for $R(u)$ is simplified to

$$R(u) = \mathbf{M}\{\xi(u) \xi(0)\}$$

In correlation theory a *stationary stochastic process* is called *continuous* if

$$\mathbf{M}(\xi(t+u) - \xi(t))^2 \rightarrow 0$$

* By this term is meant the compatibility of all distribution functions of the process.

as $u \rightarrow 0$. As follows from Chebyshev's inequality, for a continuous process, given any $\varepsilon > 0$ and arbitrary t , we have in particular the following relation:

$$P \{ |\xi(t+u) - \xi(t)| \geq \varepsilon \} \rightarrow 0 \quad (u \rightarrow 0)$$

As follows from the equation

$$M(\xi(t+u) - \xi(t))^2 = 2(1 - R(u))$$

for a continuous stationary process the relation

$$\lim_{u \rightarrow 0} R(u) = 1$$

holds.

In the case of a continuous stationary process, $R(u)$ is a continuous function of u . In fact,

$$\begin{aligned} |R(u + \Delta u) - R(u)| &= |M\{\xi(u + \Delta u)\xi(0)\} - M\{\xi(u)\xi(0)\}| = \\ &= |M\{\xi(0)[\xi(u + \Delta u) - \xi(u)]\}| \end{aligned}$$

But by the Cauchy-Bunyakovsky-Schwarz inequality,

$$|M\{\xi(0)[\xi(u + \Delta u) - \xi(u)]\}| \leq \sqrt{M\xi^2(0)M[\xi(u + \Delta u) - \xi(u)]^2}$$

And since

$$M\xi^2(0) = 1$$

and for a continuous process, as $\Delta u \rightarrow 0$,

$$M[\xi(u + \Delta u) - \xi(u)]^2 \rightarrow 0$$

it follows that, as $\Delta u \rightarrow 0$, $|R(u + \Delta u) - R(u)|$ also tends to 0. This inequality proves our assertion.

In the theorem that will now be proved, stationarity may be understood both in the broad meaning and in the narrow sense.

Khinchin's Theorem. *For a function $R(u)$ to be the correlation function of a continuous stationary process, it is necessary and sufficient that it be representable in the form*

$$R(u) = \int \cos ux dF(x) \quad (3)$$

where $F(x)$ is some distribution function.

Proof. The condition of the theorem is necessary. Indeed, if $R(u)$ is the correlation function of a continuous stationary process, then it is continuous and bounded. Let us prove, in addition, that it is positive definite. Indeed, no matter what the real numbers u_1, u_2, \dots, u_n , the complex numbers $\eta_1, \eta_2, \dots, \eta_n$ and

the integer n , the following relation holds:

$$0 \leq \mathbf{M} \left| \sum_{k=1}^n \eta_k \xi(u_k) \right|^2 = \mathbf{M} \left\{ \sum_{i=1}^n \sum_{j=1}^n \eta_i \bar{\eta}_j \xi(u_i) \xi(u_j) \right\} = \\ = \sum_{j=1}^n \sum_{i=1}^n R(u_i - u_j) \eta_i \bar{\eta}_j$$

By the Bochner-Khinchin theorem (Sec. 39), it follows from this that $R(u)$ may be represented as

$$R(u) = \int e^{iux} dF(x)$$

where $F(x)$ is a nondecreasing function of bounded variation. Whence, by virtue of the fact that the function $R(u)$ is real, we get

$$R(u) = \int \cos ux dF(x) *$$

Finally, taking into account the condition of continuity of the process,

$$R(+0) = 1$$

we find that

$$F(+\infty) - F(-\infty) = 1$$

or that $F(x)$ is some distribution function.

The condition is sufficient. We are given that $R(u)$ is a function of the form (3). We have to prove that there exists a stationary process $\xi(t)$ whose correlation function is the function $R(u)$. For this purpose, for every integral n and every group of real numbers t_1, t_2, \dots, t_n we consider the n -dimensional vector

$$\xi(t_1), \xi(t_2), \dots, \xi(t_n)$$

which is normally distributed and has the properties

$$\mathbf{M}\xi(t_1) = \mathbf{M}\xi(t_2) = \dots = \mathbf{M}\xi(t_n) = 0$$

$$\mathbf{D}\xi(t_1) = \mathbf{D}\xi(t_2) = \dots = \mathbf{D}\xi(t_n) = 1$$

For any i and j the correlation coefficient of $\xi(t_i)$ and $\xi(t_j)$ is equal to $R(t_i - t_j)$, that is,

$$\mathbf{M}\xi(t_i) \xi(t_j) = R(t_i - t_j)$$

The form of the function $R(u)$ ensures the positive definiteness of the quadratic form in the exponent of the n -dimensional normal law. The *normal* stochastic process thus defined is stationary both in the strict and the broad sense of the word.

*In consequence of the result of Example 3, Sec. 36, the function $F(x)$ is symmetric, that is,

$$F(x + 0) = 1 - F(-x)$$

This theorem plays a fundamental role in the theory of stationary processes and in its applications in physics. For details, the reader should study the specialized literature.

Example 1. Let

$$\xi(t) = \xi \cos \lambda t + \eta \sin \lambda t$$

where ξ and η are uncorrelated* random variables for which $M\xi = M\eta = 0$, $D\xi = D\eta = 1$, and λ is a constant.

Since

$$\begin{aligned} R(u) &= M\xi(t+u)\xi(t) = \\ &= M[\xi \cos \lambda(t+u) + \eta \sin \lambda(t+u)] \cdot [\xi \cos \lambda t + \eta \sin \lambda t] = \\ &= M[\xi^2 \cos \lambda t \cdot \cos \lambda(t+u) + \xi\eta(\sin \lambda(t+u) \cdot \cos \lambda t + \\ &\quad + \cos \lambda(t+u) \sin \lambda t) + \eta^2 \sin \lambda t \sin \lambda(t+u)] = \\ &= \cos \lambda t \cdot \cos \lambda(t+u) + \sin \lambda t \cdot \sin \lambda(t+u) = \cos \lambda u \end{aligned}$$

it follows that the process $\xi(t)$ is stationary in the broad sense. For this case, we have to put, in formula (3),

$$F(x) = \begin{cases} 0 & \text{for } x \leq -\lambda \\ 1/2 & \text{for } -\lambda < x \leq \lambda \\ 1 & \text{for } x > \lambda \end{cases}$$

Example 2. Let

$$\xi(t) = \sum_{k=1}^n b_k \xi_k(t)$$

where $\xi_k(t) = \xi_k \cos \lambda_k t + \eta_k \sin \lambda_k t$, λ_k are constants, $\sum_{k=1}^n b_k^2 = 1$ and the random variables ξ_k and η_k satisfy the following conditions:

$$\begin{aligned} M\xi_k &= M\eta_k = 0, \quad D\xi_k = D\eta_k = 1 \quad (k = 1, 2, \dots, n) \\ M\xi_i \xi_j &= M\eta_i \eta_j = 0 \quad \text{for } i \neq j \\ M\xi_i \eta_j &= 0 \quad \text{for } i, j = 1, \dots, n \end{aligned}$$

It is easy to compute that the correlation function for $\xi(t)$ is equal to

$$R(u) = \sum_{k=1}^n b_k^2 \cos \lambda_k u$$

and that, consequently, the process $\xi(t)$ is a stationary process in the broad sense of the word. The function $F(x)$ in formula (3)

* The random variables ξ and η are termed *uncorrelated* if $M\xi\eta = M\xi \cdot M\eta$.

only increases at the points $\pm\lambda_k$ and has jumps of magnitude $\frac{1}{2} b_k^2$ at these points.

Stochastic processes for which the function $F(x)$ increases by jumps alone are called *processes with discrete spectra*.

It is easy to see that any process of the type

$$\xi(t) = \sum_{k=1}^{\infty} b_k \xi_k(t) \quad (4)$$

where $\sum_1^{\infty} b_k^2 < \infty$ and $\xi_k(t)$ have the same meaning as in Example 2 is stationary in the broad sense and has a discrete spectrum. It is important to point out that E. E. Slutsky found a profound converse proposition: *Any stationary process with a discrete spectrum is representable in the form of (4)*. Generalization of Slutsky's theorem to the case of an arbitrary spectrum will be formulated in the next section.

In parallel with the theory of stationary processes, there developed a theory of stationary sequences. A sequence of random variables

$$\dots, \xi_{-2}, \xi_{-1}, \xi_0, \xi_1, \xi_2, \dots \quad (5)$$

is called *stationary* if for any integral n, u and t_j ($1 \leq j \leq n$) condition (1) is satisfied. Similarly, sequence (5) is called *stationary in the broad sense* if for all terms of the sequence the expectations and variances are constant numbers that are independent of their place in the sequence

$$\begin{aligned} \dots &= M\xi_{-2} = M\xi_{-1} = M\xi_0 = M\xi_1 = M\xi_2 = \dots = a \\ \dots &= D\xi_{-2} = D\xi_{-1} = D\xi_0 = D\xi_1 = D\xi_2 = \dots = \sigma^2 \end{aligned}$$

and the correlation coefficient of ξ_i and ξ_j is a function solely of $|i-j|$.

By way of an exercise, we suggest that the reader prove the theorem: if for a stationary sequence

$$\lim_{s \rightarrow \infty} R(s) = 0$$

where $R(s)$ is the correlation coefficient of ξ_i and ξ_{i+s} , then the law of large numbers applies; that is, as $n \rightarrow \infty$,

$$P \left\{ \left| \frac{1}{n} \sum_{k=1}^n \xi_k - a \right| < \varepsilon \right\} \rightarrow 1$$

no matter what the constant $\varepsilon > 0$.

*Sec. 58. The Concept of a Stochastic Integral.
The Spectral Decomposition of Stationary Processes*

For what follows we have to introduce the concept of a stochastic integral. Let a stochastic process $\xi(t)$ and a numerical function $f(t)$ be given on the interval $a \leq t \leq b$. Partition the interval $[a, b]$ by points $a = t_0 < t_1 < \dots < t_n = b$ and consider the sum

$$I_n = \sum_{i=1}^n f(t_i) \xi(t_i) (t_i - t_{i-1})$$

If as $\max_{1 \leq i \leq n} (t_i - t_{i-1}) \rightarrow 0$ this sum tends to a certain limit (which, generally speaking, is a random variable), then the limit is called the *integral of the stochastic process $\xi(t)$* and is denoted by the symbol

$$I = \int_a^b f(t) \xi(t) dt$$

The improper integral (for $a = -\infty$, $b = \infty$) is defined in the usual manner as the limit of proper integrals as $a \rightarrow -\infty$, $b \rightarrow \infty$.

The convergence of the integral sums I_n is to be understood in the following sense: there exists a random variable I such that as $n \rightarrow \infty$

$$\mathbf{M} (I_n - I)^2 \rightarrow 0 \quad (1)$$

Proceeding from familiar theorems in the theory of functions of a real variable, it is easy to prove that the sequence of random variables I_n converges to the limit I in the sense of (1) if and only if

$$\mathbf{M} (I_m - I_n)^2 \rightarrow 0 \quad (2)$$

as $\min(m, n) \rightarrow \infty$. We shall not dwell on the proof of this fact.

Theorem 1. *For the integral*

$$I = \int_a^b f(t) \xi(t) dt$$

to exist it is sufficient that the integral

$$A = \int_a^b \int_a^b R(t-s) f(t) f(s) ds dt$$

should exist. Moreover,

$$A = \mathbf{M} \left[\int_a^b f(t) \xi(t) dt \right]^2$$

Proof. To prove the first half of the theorem it will suffice to notice that if the integral A exists, then the relation (2) holds. And we have

$$\begin{aligned}
 \mathbf{M}(I_n - I_m)^2 &= \\
 &= \mathbf{M} \left[\sum_{i=1}^n f(t_i) \xi(t_i) \Delta t_i \right]^2 - 2 \mathbf{M} \sum_{i=1}^n \sum_{j=1}^m f(t_i) f(s_j) \xi(t_i) \xi(s_j) \Delta t_i \Delta s_j + \\
 &\quad + \mathbf{M} \left[\sum_{j=1}^m f(s_j) \xi(s_j) \Delta s_j \right]^2 = \\
 &= \sum_{i=1}^n \sum_{k=1}^n f(t_i) f(\tau_k) R(t_i - t_k) \Delta t_i \Delta \tau_k - \\
 &\quad - 2 \sum_{i=1}^n \sum_{j=1}^m f(t_i) f(s_j) R(t_i - s_j) \Delta t_i \Delta s_j + \\
 &\quad + \sum_{j=1}^m \sum_{k=1}^m f(s_j) f(\sigma_k) R(s_j - \sigma_k) \Delta s_j \Delta \sigma_k
 \end{aligned}$$

Here, the numerical values of t_k and τ_k and also s_k and σ_k coincide.

By virtue of the assumption that the integral A exists,

$$\begin{aligned}
 A &= \lim \sum_{i=1}^n \sum_{k=1}^n f(t_i) f(\tau_k) R(t_i - t_k) \Delta t_i \Delta \tau_k = \\
 &= \lim \sum_{i=1}^n \sum_{j=1}^m f(t_i) f(s_j) R(t_i - s_j) \Delta t_i \Delta s_j = \\
 &= \lim \sum_{j=1}^m \sum_{k=1}^m f(s_j) f(\sigma_k) R(s_j - \sigma_k) \Delta s_j \Delta \sigma_k
 \end{aligned}$$

so long as $\max[\Delta t_i, \Delta s_j] \rightarrow 0$. Thus, as $\min(m, n) \rightarrow \infty$,

$$\mathbf{M}(I_m - I_n)^2 \rightarrow 0$$

To prove the second part of the theorem, note that

$$\begin{aligned}
 \mathbf{M} \left[\sum_{i=1}^n f(t_i) \xi(t_i) \Delta t_i \right]^2 &= \mathbf{M} \sum_{j=1}^n \sum_{i=1}^n f(t_i) f(\tau_j) \xi(t_i) \xi(\tau_j) \Delta t_i \Delta \tau_j = \\
 &= \sum_{j=1}^n \sum_{i=1}^n f(t_i) f(\tau_j) R(t_i - \tau_j) \Delta t_i \Delta \tau_j
 \end{aligned}$$

As $\max_{1 \leq i \leq n} \Delta t_i \rightarrow 0$, the last sum tends to the integral A .

In addition to the notion of a stochastic integral that has just been introduced, we can also consider a stochastic Stieltjes integral, which we define as the limit of the sums

$$\sum_{k=1}^n f(t_k) [\xi(t_k) - \xi(t_{k-1})] \quad (3)$$

as $\max(t_i - t_{i-1}) \rightarrow 0$. As before, $a = t_0 \leq \dots \leq t_n = b$ and the limit is to be understood in the sense of (1). If the limit of the sums (3) exists, we will denote it by the symbol

$$\int_a^b f(t) d\xi(t)$$

At the conclusion of Sec. 57 we formulated Slutsky's theorem, which expresses the relationship between stationary processes with discrete spectra and Fourier series with random uncorrelated coefficients. It may be proved that the following property holds for every stationary process (in the broad sense): *For any $\varepsilon > 0$ and arbitrarily large T , there exist pairwise uncorrelated random variables $\xi_1, \xi_2, \dots, \xi_n, \eta_1, \eta_2, \dots, \eta_n$ and real numbers $\lambda_1, \lambda_2, \dots, \lambda_n$ * such that for any t of the interval $-T \leq t \leq T$, the inequality*

$$\mathbf{M} \left[\xi(t) - \sum_{k=1}^n (\xi_k \cos \lambda_k t + \eta_k \sin \lambda_k t) \right]^2 < \varepsilon$$

holds. From this, in particular, it follows that under the given conditions

$$\mathbf{P} \left\{ \left| \xi(t) - \sum_{k=1}^n (\xi_k \cos \lambda_k t + \eta_k \sin \lambda_k t) \right| > \eta \right\} \leq \frac{\varepsilon}{\eta^2}$$

where η is a preassigned positive number.

The following important theorem is given without proof.

Theorem 2. *Any stochastic process that is stationary in the broad sense is representable in the form*

$$\xi(t) = \int_0^\infty \cos \lambda t dZ_1(\lambda) + \int_0^\infty \sin \lambda t dZ_2(\lambda) \quad (4)$$

where the stochastic processes $Z_1(\lambda)$ and $Z_2(\lambda)$ (the variable $\lambda \geq 0$) possess the following properties:

(a) $\mathbf{M} [Z_i(\lambda_1 + \Delta\lambda_1) - Z_i(\lambda_1)] [Z_j(\lambda_2 + \Delta\lambda_2) - Z_j(\lambda_2)] = 0$ ($i, j = 1, 2$)

if $i \neq j$ and if the intervals $(\lambda_1, \lambda_1 + \Delta\lambda_1)$ and $(\lambda_2, \lambda_2 + \Delta\lambda_2)$ are nonoverlapping, then i and j may be equal;

(b) $\mathbf{M} [Z_1(\lambda + \Delta\lambda) - Z_1(\lambda)]^2 = \mathbf{M} [Z_2(\lambda + \Delta\lambda) - Z_2(\lambda)]^2$

* The numbers n and $\lambda_1, \lambda_2, \dots, \lambda_n$ and also the variables ξ_i and η_i are dependent on ε and T .

It is natural to call formula (4) the *spectral decomposition of the process* $\xi(t)$.

The stochastic processes $Z_1(\lambda)$ and $Z_2(\lambda)$ of formula (4) may be determined by the equalities

$$Z_1(\lambda) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{\sin \lambda t}{t} \xi(t) dt$$

and

$$Z_2(\lambda) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{1 - \cos \lambda t}{t} \xi(t) dt$$

It is easy to prove that both integrals exist (this is done by means of Khinchin's formula that was proved in Sec. 57). It will also be seen that

$$F(\lambda + \Delta\lambda) - F(\lambda) = M [Z_1(\lambda + \Delta\lambda) - Z_1(\lambda)]^2$$

where the function $F(\lambda)$ is determined by Khinchin's theorem.

The possibility of decomposing into the form (4) an arbitrary stochastic process which is stationary in the broad sense was pointed out in 1940 by A. N. Kolmogorov, who stated the result in terms of Hilbert space geometry and proved it by means of the spectral theory of operators. Since then, many authors, such as Cramér, Karhunen, Loève, Blanc-Lapierre and others, have contributed to the probabilistic interpretation and derivation of this decomposition.

We shall not speak here of the applications of the spectral decomposition to problems in the theory of oscillations and geophysics; we refer the reader to the works of A. M. Yaglom, Blanc-Lapierre and Fortet, which are given in the bibliography at the end of the book.

Sec. 59. The Birkhoff-Khinchin Ergodic Theorem

In 1931 the American mathematician George Birkhoff proved a general theorem of mechanics, which, as A. Ya. Khinchin demonstrated three years later, allows for a broad probabilistic generalization. The theorem is as follows: *if a continuous stationary process $\xi(t)$ has a finite expectation, then with probability one there exists the limit*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \xi(t) dt$$

The stationarity of the process is here understood in the strict and not broad sense.

Since this theorem is a peculiar form of the strong law of large numbers, we will prove it for stationary sequences (not for stationary processes) so as to continue directly the formulations of Chapter 6.

Theorem. *For a stationary sequence of random variables*

$$\dots, \xi_{-1}, \xi_0, \xi_1, \dots$$

for which $M\xi_j$ is finite, the sequence of arithmetic means

$$\frac{1}{n} \sum_{i=1}^n \xi_i$$

converges to a limit with probability one.

Proof. We denote

$$h_{ab} = \frac{\xi_a + \xi_{a+1} + \dots + \xi_{b-1}}{b-a}$$

We have to prove that with probability one the quantities h_{0b} tend to a limit as $b \rightarrow \infty$. Denote the random event that this limit exists by the letter \bar{K} . We must prove that $P(\bar{K})=1$ or, which is the same thing, that $P(K)=0$.

We assume the contrary, that the event K (that is, that the quantities h_{0b} do not converge to a limit as $b \rightarrow \infty$) has a positive probability and we will demonstrate that this assumption leads to a contradiction.

For this purpose we consider all intervals (α_n, β_n) with rational endpoints, $\alpha_n < \beta_n$. The set of all such intervals is countable. If $\lim_{b \rightarrow \infty} h_{0b}$ does not exist, there will be an interval (α_n, β_n) for which $\limsup_{b \rightarrow \infty} h_{0b} > \beta_n$ and $\liminf_{b \rightarrow \infty} h_{0b} < \alpha_n$ (event K_n). Thus, the event K decomposes into a countable set of mutually exclusive cases K_n . Since by assumption $P(K) > 0$, an n can be found such that $P(K_n) > 0$.

It is thus proven that if $P(K) > 0$, there exist two numbers α and β ($\alpha < \beta$) for which the following inequalities hold simultaneously:

$$\left. \begin{aligned} \limsup h_{0b} &> \beta \\ \liminf h_{0b} &< \alpha \end{aligned} \right\} \quad (1)$$

Now suppose that all ξ_j have taken on certain definite values. If the interval (a, b) is such that $h_{ab} > \beta$, but for all b' , for which $b < b' < b$, $h_{ab'} \leq \beta$, then this interval will be called *special* (with respect to β).

It will readily be seen that two special intervals do not overlap. Indeed, if two special intervals (a, b) and (a_1, b_1) are such that $a <$

$a_1 < b < b_1$, then from the equality

$$h_{ab} = \frac{(a_1 - a) h_{aa_1} + (b - a_1) h_{a_1b}}{b - a}$$

and the inequality $h_{ab} > \beta$ there follows either $h_{aa_1} > \beta$ or $h_{a_1b} > \beta$. However, the first of these inequalities is impossible because the interval (a, b) is special, the second is also impossible since the interval (a_1, b_1) is special.

The difference $b - a$ will be called the *rank (length) of the interval* (a, b) . If an interval (a, b) is special, is of a rank exceeding s , and is not contained in any interval of a rank exceeding s , then such an interval will be called *s-special*.

Since from among special intervals containing an arbitrary interval (α, β) of length not exceeding s and also having a length that does not exceed s , there should be at least one of greatest length, if there were two they would overlap, which is impossible on the basis of what has already been proved. Thus, every special interval of length not exceeding s may lie inside only one *s-special* (or coincide with it). From the definition it follows that two *s-special* intervals can only lie one outside the other.

We denote by K_s the event that inequalities (1) are valid and, besides, that there exists a $t \leq s$ such that $h_{0t} > \beta$. Since K is the limit for the events K_s ,

$$P(K) = \lim_{s \rightarrow \infty} P(K_s)$$

From this it follows that for all sufficiently large s the inequality $P(K_s) > 0$ is valid. From now on we will confine ourselves solely to such values of s .

Let the event K_s take place. Then among the values of those $t \leq s$ for which $h_{0t} > \beta$ there exists a least t' . The interval $(0, t')$ is special. Consequently, it lies in some *s-special* interval (a, b) (or is such itself), for which $a \leq 0 < b$. The converse is also true: if there exists an *s-special* interval (a, b) for which $a \leq 0 < b$, then there exists a $t \leq s$ such that $h_{0t} > \beta$. For $a=0$ this is obvious: it suffices to put $t=b$. But if $a < 0$, then from the equality

$$h_{ab} = \frac{-ah_{a0} + bh_{0b}}{b-a}$$

and the inequalities $h_{ab} > \beta$, $h_{a0} \leq \beta$ there follows $h_{0b} > \beta$. Thus, in this case too it is possible to set $t=b$.

We denote $-a$ by p and $b-a$ by q . Since the *s-special* interval $(-p, -p+q)$ can only exist alone, the event K_s is decomposed into the mutually exclusive cases K_{pq} , which correspond to the existence of *s-special* intervals $(-p, -p+q)$:

$$K_s = \sum_{p, q} K_{pq} \quad (q = 1, \dots, s, p = 0, 1, \dots, q-1)$$

Changing the numbering of the sequence $i' = i + p$ transforms the case K_{0q} into the case K_{pq} . Therefore, by virtue of stationarity*,

$$\begin{aligned} P(K_{pq}) &= P(K_{0q}) \text{ and } M\xi_{0q}/K_{pq} = M\xi_p/K_{0q}. \text{ Since} \\ P(K_s) M\xi_{0q}/K_s &= \sum_{p, q} P(K_{pq}) M\xi_{0q}/K_{pq} = \sum_q P(K_{0q}) \sum_p M\xi_p/K_{0q} = \\ &= \sum_q P(K_{0q}) Mqh_{0q}/K_{0q} \end{aligned}$$

we find, by taking into account that in the event of K_{0q} the inequality $h_{0q} > \beta$ holds, that

$$P(K_s) M\xi_{0q}/K_s > \sum_q P(K_{0q}) q\beta = \beta \sum_{p, q} P(K_{pq}) = \beta P(K_s)$$

Whence, since by assumption $P(K_s) \neq 0$, it follows that

$$M\xi_{0q}/K_s > \beta$$

Since $K_s \rightarrow K$, we have

$$M\xi_{0q}/K \geq \beta$$

In similar fashion (if special intervals were considered with respect to α) it is possible to prove that

$$M\xi_{0q}/K \leq \alpha$$

This is a contradiction. And so it follows that $P(K) = 0$, which is what we set out to prove.

To investigate the limit to which the quantities h_{0n} tend as $n \rightarrow \infty$ requires additional arguments. We confine ourselves here to the proof of the following theorem.

Theorem. *If the random variables ξ_k are stationary, have finite variance and the correlation function $R(k) \rightarrow 0$ as $k \rightarrow \infty$, then*

$$P\left\{h_{0n} \rightarrow a\right\} = 1 \quad (a = M\xi_k)$$

Proof. Consider the variance of the quantity h_{0n} . By virtue of stationarity we have

$$Dh_{0n} = M\left[\frac{1}{n} \sum_{k=1}^n (\xi_k - a)\right]^2 = \frac{D\xi_n}{n^2} \left[n + 2 \sum_{1 \leq i < j \leq n} R(j-i)\right]$$

It is obvious that

$$\sum_{1 \leq i < j \leq n} R(j-i) = \sum_{k=1}^{n-1} (n-k) R(k)$$

* Note that only at this point have we made use of the assumption of stationarity.

Consider an m so large that for $k > m$ we have the inequality

$$|R(k)| \leq \varepsilon \quad (\varepsilon > 0)$$

From this it follows that

$$Dh_{0n} \leq \frac{D\xi_n}{n^2} \left[n + 2 \sum_{k=1}^m (n-k) R(k) + 2\varepsilon \sum_{k=m+1}^{n-1} (n-k) \right]$$

This inequality is obviously strengthened as follows:

$$Dh_{0n} \leq \frac{D\xi_n}{n^2} [n + 2m(n-1) + \varepsilon(n-m-1)(n-m)]$$

From this it is clear that if n is sufficiently great, the right side of this inequality may be made less than 3ε . Thus, as $n \rightarrow \infty$ the quantities h_{0n} converge in probability to a , and since h_{0n} converge with probability one, as n tends to infinity, the assertion of the theorem is obviously correct.

The above-proved theorem is not only of considerable theoretical interest, but finds extensive applications in statistical physics and in engineering practice as well. The reason is that to determine such important characteristics of a phenomenon as $M\xi(t)$, $D\xi(t)$, $R(u)$ in the case of stationary processes, one does not need to know the probability distribution of the possible values or to calculate these quantities from appropriate formulas. The determination of these *spatial averages*, to use the physics term, demands of the investigator information that is often lacking. At any rate, the experimental estimation of these quantities requires repeated realization of trials for the process [that is, numerous realizations of the function $\xi(t)$ have to be obtained from experiment]. The Birkhoff-Khinchin ergodic theorem shows that it is possible, with probability one, to confine oneself (under specific conditions) to a single realization of the process $\xi(t)$.

Elements of Queueing Theory

Sec. 60. A General Description of the Problems of the Theory

Of the numerous and profound applications of the theory of stochastic processes to various problems of physics, biology, engineering and economics we consider here only one, which in recent years has seen considerable development under the impetus of the diversified demands of practice. Originally, the specific problems that lead to the theory of mass-scale service, as it is termed in the Soviet literature (or of the queueing process), emerged in connection with the operation of telephone systems. Later it was found that similar problems arise in merchandizing (computing the number of shops, salesmen, cash-registers, supplies, etc.), in the operation of production equipment, in calculating the traffic capacity of roads, bridges, crossings, aerodromes, canal locks, seaports, and so forth. A. K. Erlang, a Danish scientist engaged for many years in the Copenhagen Telephone Company, played the basic role in formulating and solving the first mathematical problems of this nature. Today they constitute but a small fraction of the problems of this nature that have been elaborated. Interest in queueing theory shot up throughout the world and the number of theoretical and applied studies in this field far exceeds a thousand.

To get an idea of the peculiarities of stating problems in queueing theory, we shall first examine a few problems of an applied nature and remain on a purely qualitative level. We will then take one problem (given the most elementary assumptions) and will study the classical methods of Erlang. The sequel will be clear from the table of contents.

Suppose that a telephone exchange receives calls from its subscribers. If at the time of arrival of a call there are open lines, the subscriber is switched to one of them and a conversation is started that lasts as long as is necessary. If all lines are engaged, various systems of servicing the subscriber are possible. At the present time, two systems have been worked out in detail: a waiting system and a system involving

losses. In the former, a call that arrives at the exchange when all lines are busy is put in a waiting line and is made to wait until all the calls that arrived earlier have been completed. In the latter system, a call arriving at the exchange when all lines are engaged gets a refusal, and all subsequent servicing proceeds as if that call had not arrived (we say that this is "loss of a call").

For us at this point it is important to stress two peculiarities that must be taken into account when considering the problems that arise. The first is that the calls arrive at the telephone exchange at random instants of time and it is not possible to predict in advance when the next call will be made. Similarly, the duration of talks is not a constant factor and varies at random. Later on we shall return to a more detailed consideration of these two peculiarities that are found in all problems of queueing theory.

The servicing systems involving waiting and losses do not only differ as to engineering aspects of the devices that handle them but also in the mathematical problems that emerge in their study. Indeed, to estimate the quality of servicing in the waiting system it is particularly essential to determine the mean waiting time for the start of service, that is, the mean waiting (sojourn) time in a queue. For systems involving losses, the waiting time is of no interest at all, whether engineering or mathematical. Here the important thing is the probability of congestion (loss of a call). But whereas in the second servicing system the probability of congestion affords a sufficiently complete picture of what may be expected in the given setup, the situation is more complicated in the waiting system. Despite its importance, the mean waiting time is not an exhaustive characteristic of the quality of service. Another very important factor is the spread of the waiting time about the mean. Also of interest is the distribution of the length of the waiting line, the extent of the load on the service equipment, the distribution of duration of continuous operation of the equipment.

The situation at a ticket office with a waiting line is similar to the system of servicing subscribers at a telephone exchange with queues. Some large factories have tool-supply depots. If such a depot services a large number of workmen, skilled workers find that they lose time waiting; if there are more depots, the supply clerks are idle much of the time. A similar problem arises in organizing the work of a seaport. Cargo ships do not arrive in port according to a fixed schedule, and loading and unloading times vary. If there is a lack of docking facilities, the ships spend appreciable times waiting, and this represents a loss, economically speaking. But surplus docking facilities increase the idle time of equipment and workmen. This gives rise to an important economic problem of determining the optimal range of docking facilities for handling a given freight turnover at minimal loss involved in maintenance of ships and handling equipment.

During the 1930s, in connection with expanding automatization of machine-tools in industry, there was a trend towards one operator servicing a number of machines. At random times, for various reasons, the machines go out of commission and require the attention of the workman (repairman). The duration of the repair operation is, generally speaking, not constant and is a random variable. The question then arises: How great is the probability that at a given instant of time a certain number of machines will be waiting to be serviced? Important practical questions follow: What is the mean idling time of the machines in charge of one workman? For a given setup, how many machines can one workman handle most efficiently?

Many problems of a scientific, manufacturing and economic nature involve more than just systems with waiting and losses. In everyday affairs we ourselves know how frequently one has to refuse service because of long waiting times. For example, in interurban (long-distance) telephone calls we often have to limit the waiting time and warn the operator that if the connection is not made within a specified time, the call is to be cancelled. The problem is much the same in the sale of perishable food, in the organization of medical aid, the operation of a large airport, and so forth.

It is therefore quite natural to formulate the following group of related problems. A service system receives certain requests. If there is at least one free server (servicing device), the incoming request is handled immediately. If all the servers are busy, a fresh request gets in line: (a) if there are no more than a given number m of requests; (b) for a length of time not greater than τ (this time is constant or depends on chance); (c) for as long as is necessary but is serviced during a time not greater than τ (at the expiration of this time the request leaves the waiting line even if not completely serviced); (d) but in such manner that the sojourn time in the system (the total waiting time and servicing time) does not exceed τ .

In the foregoing problems, we proceeded on the assumption that the servicing devices (servers) were absolutely reliable and were constantly in operation. It is of course far removed from any actual situation. There naturally arises the important problem of taking into account the effect of malfunctioning (breakdowns) of servers on the effectiveness of a service system.

From now on we shall speak of demands made on servers for service. The totality of moments at which demands for service arise constitute a stochastic process. We will call this process the *incoming flow of demands (incoming traffic)*. The incoming traffic may be described by the process $k(t)$, which signifies the number of demands that arrive between time 0 and time t . In the overwhelming majority of papers dealing with queueing theory it is assumed that the incoming traffic constitutes a *Poisson process* (or, as it is sometimes called, an *elementary flow*) which was described in Sec. 51. There, the conditions were

given under which an elementary flow occurs. Later on, in Sec. 63, we will give other conditions that will ensure an elementary flow.

The service time is a random variable with a certain distribution function $H(x)$. It is very often considered, in investigations of both a theoretical and applied nature, that $H(x)=0$ for $x \leq 0$ and $H(x)=1-e^{-\nu x}$ for $x > 0$, where ν is a positive constant. This choice is not by chance but is due to a number of circumstances associated with simplicity of solution (this will be discussed later on). For the present we confine ourselves to the proof of an important property of exponential distribution.

Theorem. *If the service time has an exponential distribution: for $x > 0$*

$$H(x)=1-e^{-\nu x} \quad (1)$$

where ν is a positive constant, then the distribution of the remaining part of the service time does not depend on how long the servicing has been in progress.

Proof. Let $h_a(t)$ denote the probability that the servicing, which has already continued for a time a , will continue for at least a time t . From assumption (1) it is clear that

$$h_0(a+t)=e^{-(a+t)\nu}, \quad h_0(a)=e^{-\nu a}$$

Since by the multiplication theorem

$$h_0(a+t)=h_0(a)h_a(t)$$

it follows that

$$e^{-\nu(a+t)}=e^{-\nu a}h_a(t)$$

From this we get

$$h_a(t)=e^{-\nu t}$$

and the proof is complete.

To illustrate problems of queueing theory let us examine service with loss under conditions that were studied by Erlang.

There are n servers to which there is an incoming flow of elementary demands. Every device (server) is accessible to any demand when it is free. Every demand is serviced by one server only, and every server serves only one demand (when it is busy). A demand that finds all servers busy servicing other demands is lost. Our problem consists in finding the probability of congestion.

Let us consider the process of change of state of our system of service. At each instant of time it can be in one of the following states: E_0 signifies that all servers are free, E_1 , one server is busy, the remaining are free, ..., E_n means that all servers (devices) are busy. Let us see what peculiarities this process has under the assumptions that we have made.

At some time t_0 let our system be in state E_k . We shall prove that the subsequent course of the process is fully determined by this and does not depend on what occurred prior to time t_0 . In other words, the process under consideration is a Markov process. Indeed, the subsequent course of the process is determined completely by the following three factors:

- (1) the times at which the servicings that are accomplished at t_0 terminate;
- (2) the times at which new demands appear;
- (3) the duration of service of demands that appear in the system after t_0 .

By the peculiarity (just proven) of an exponential distribution, the durations of the remaining parts of service are not dependent on how long the service continued prior to t_0 . Since the traffic (flow of demands) is elementary, the past has no effect on how many demands arrive after t_0 . Finally, the service time of demands appearing after t_0 is in no way dependent on what was serviced (and how) prior to this moment. This shows that the process of variation of the system under study is Markovian. This circumstance is fundamental inasmuch as it permits obtaining simple equations for those characteristics of the process that interest us.

Denote by $p_k(t)$ the probability that at time t the system is in state E_k . Form the equations for the functions $p_k(t)$.

First we find the probability that at time $t+h$ all the servers are free. This can occur in the following mutually exclusive ways: at time t all servers were free and during time h no new demands arrived; at time t one server was busy, servicing was terminated during the subsequent time interval h , and no new demands arrived; the remaining possibilities—two servers were busy and during time h they completed their servicing, etc.—have probability of the order of $o(h)$. The probability of the first event is

$$p_0(t) e^{-\lambda h} = p_0(t) (1 - \lambda h + o(h))$$

the probability of the second event is

$$p_1(t) e^{-\lambda h} (1 - e^{-\nu h}) = p_1(t) \nu h + o(h)$$

Thus,

$$p_0(t+h) = p_0(t) (1 - \lambda h) + \nu h p_1(t) + o(h)$$

whence, by passing to the limit as $h \rightarrow 0$, we get the following equation:

$$p'_0(t) = -\lambda p_0(t) + \nu p_1(t) \quad (2)$$

Reasoning in similar fashion for $1 \leq k < n$ we get

$$p'_k(t) = \lambda p_{k-1}(t) - (\lambda + k\nu) p_k(t) + (k+1) \nu p_{k+1}(t) \quad (3)$$

and for $k = n$

$$p'_n(t) = \lambda p_{n-1}(t) - n\nu p_n(t) \quad (4)$$

The system we have obtained of linear differential equations permits us to find the desired functions $p_k(t)$. The arbitrary constants are determined by means of initial data, which we choose as follows:

$$p_0(0) = 1, \quad p_k(0) = 0 \text{ for } k \geq 1$$

These equations mean that at the initial instant all servers were free. We add that the probabilities $p_k(t)$ must satisfy yet another additional normalizing condition:

$$\sum_{k=0}^n p_k(t) = 1 \quad (5)$$

Interest ordinarily centres on studying a steady-state process, that is, the solution is considered as $t \rightarrow \infty$. As we shall see in the next section, under the conditions of our problem there are the limits

$$p_k = \lim_{t \rightarrow \infty} p_k(t)$$

and these limiting probabilities satisfy the following set of algebraic equations obtained from (2)-(5) by substituting the constants p_k for the functions $p_k(t)$ and zeros for the derivatives $p'_k(t)$:

$$\begin{aligned} -\lambda p_0 + \nu p_1 &= 0 \\ \lambda p_{k-1} - (\lambda + k\nu) p_k + (k+1)\nu p_{k+1} &= 0 \quad (1 \leq k < n) \\ \lambda p_{n-1} - n\nu p_n &= 0 \\ \sum_{k=0}^n p_k &= 1 \end{aligned} \quad (6)$$

By putting $z_k = \lambda p_{k-1} - k\nu p_k$ we reduce the set of our equations to

$$z_1 = 0, \quad z_k - z_{k+1} = 0 \text{ for } 1 \leq k < n, \quad z_n = 0$$

whence we find that

$$k\nu p_k = \lambda p_{k-1}, \quad k = 1, 2, \dots, n$$

Simple transformations result in the equalities

$$p_k = \frac{\rho^k}{k!} p_0 \quad \left(k \geq 1, \rho = \frac{\lambda}{\nu} \right)$$

Now (6) enables us to find the normalizing factor p_0 :

$$p_0 = \left[\sum_{k=0}^n \frac{\rho^k}{k!} \right]^{-1}$$

Finally,

$$p_k = \frac{\rho^k}{k!} \left[\sum_{j=0}^n \frac{\rho^j}{j!} \right]^{-1} \quad (7)$$

These formulas were found by Erlang and are called *Erlang's formulas*. For $k=n$ we obtain the probability that all servers are busy and, consequently, the probability that every new demand (call) arriving in the system will be lost. Thus, the probability of rejection (congestion) is

$$p_n = \frac{\frac{1}{n!} \rho^n}{\sum_{k=0}^n \frac{1}{k!} \rho^k}$$

To illustrate the rapidity of increase in the probability of losses with increasing ρ (the load per server), we give the following tables. Here we confine ourselves solely to the cases of $n=2$ and $n=4$ and such values of ρ for which, on the average, every server has the same intensity of incoming demands.

$n = 2$

ρ	0.1	0.3	0.5	1.0	2.0	3.0	4.0
p_n	0.0045	0.0335	0.0769	0.2000	0.4000	0.5294	0.6054

$n = 4$

ρ	0.2	0.6	1.0	2.0	4.0	6.0	8.0
p_n	0.0001	0.0030	0.0154	0.0952	0.3107	0.4696	0.5746

Examining these tables we note that for small loads, an increase in the number of servers substantially reduces the probability of losses inasmuch as the probability of all servers being busy when there are a large number of servers is small. But then as the load on every server increases, the probability of losses gradually levels off.

Sec. 61. Birth and Death Processes

Erlang's problem and many other problems of queueing theory considered under the elementary assumptions spoken of at the end of the preceding section fit into a scheme that bears the name "birth and death processes". This class of processes attracted attention in connection with biological problems of the sizes of population, the spread of epidemics, and so forth. Since the mathematical scheme underlying birth and death processes is of a sufficiently general nature, the theory was broadly applied to many other problems as well.

Let us suppose that a certain system can at every instant of time be in one of the states E_0, E_1, E_2, \dots , the set of which is finite or countable. The states of the system change with time, and during an interval of duration h the system will pass from state E_n at time t to state E_{n+1} with probability $\lambda_n h + o(h)$ and to state E_{n-1} with probability $\nu_n h + o(h)$. The probabilities that during the time interval $(t, t+h)$ the system will pass to state $E_{n \pm k}$ for $k > 1$ are infinitely small in comparison with h . From this it follows that the probability of staying in the state E_n during the same time interval is equal to $1 - \lambda_n h - \nu_n h + o(h)$. The constants λ_n and ν_n are assumed to be dependent on n but independent of t and of how the system got to that state. This latter circumstance signifies that the process at hand is Markovian. The theory that will be given here can be extended to the case when the quantities λ_n and ν_n are dependent on t .

The stochastic process just described goes by the name *birth and death process*. If by E_n is understood an event that the size of a population is n , then the transition $E_n \rightarrow E_{n+1}$ means that the size of the population has increased by unity. Similarly, the transition $E_n \rightarrow E_{n-1}$ is to be regarded as the death of one member of the population.

If for any $n \geq 1$ the equalities $\nu_n = 0$ hold, that is, if only the transitions $E_n \rightarrow E_{n+1}$ are possible at the time of change of state, then the process is called a *birth process* (the phrase "a pure birth process" is sometimes used). But if all $\lambda_n = 0$ ($n = 0, 1, 2, \dots$), then the process is a *death process*.

The Poisson process that we studied in Sec. 51 is a birth process; here, $\lambda_n = \lambda$ for all $n \geq 0$.

The Erlang problem that we examined in the preceding section is also a birth and death process for which $\lambda_k = \lambda$ when $0 \leq k < n$ and $\lambda_k = 0$ for $k \geq n$; $\nu_k = 0$ for $k > n$ and $\nu_k = kv$ for $1 \leq k \leq n$.

We denote by $p_k(t)$ the probability that the system we are studying is in state E_k at time t . Reasoning like we did in Sec. 51 and when we considered the Erlang problem, we come to a system of differential equations that governs the birth and death process:

$$p'_0(t) = -\lambda_0 p_0(t) + \nu_1 p_1(t) \quad (1)$$

and for $k \geq 1$

$$p'_k(t) = -(\lambda_k + \nu_k) p_k(t) + \lambda_{k-1} p_{k-1}(t) + \nu_{k+1} p_{k+1}(t) \quad (2)$$

Our notations are somewhat deficient in that we have not indicated from what state E_i the system began to change. An exhaustive notation would be $p_{ij}(t)$ —the probability that the system will, at time t , be in the state E_j if at time 0 it was in the state E_i . In Sec. 51 and in Erlang's problem we assumed that E_0 was the initial state.

Equations (1) and (2) become especially simple for processes of pure death and pure birth. In the latter instance, after performing successive integration (the formulas are written on the assumption that all λ_k are different) we get

$$\begin{aligned} p_0(t) &= e^{-\lambda_0 t} \\ p_1(t) &= \frac{\lambda_0}{\lambda_1 - \lambda_0} (e^{-\lambda_0 t} - e^{-\lambda_1 t}) \\ p_2(t) &= \frac{\lambda_0 \lambda_1}{\lambda_1 - \lambda_0} \left[\frac{1}{\lambda_2 - \lambda_0} (e^{-\lambda_0 t} - e^{-\lambda_2 t}) + \frac{1}{\lambda_2 - \lambda_1} (e^{-\lambda_1 t} - e^{-\lambda_2 t}) \right] \end{aligned}$$

Here we assumed that for $t=0$ the system is in state E_0 . There is no difficulty in writing the general solution and seeing that the functions $p_k(t)$ are nonnegative for all k and t . However, if λ_k grow too rapidly as k increases, it may happen that $\sum_{k=0}^{\infty} p_k(t) < 1$.

Feller's Theorem. *In order that for all values of t the solutions $p_k(t)$ of the equations of pure birth satisfy the relation*

$$\sum_{k=0}^{\infty} p_k(t) = 1 \quad (3)$$

it is necessary and sufficient that the series

$$\sum_{k=0}^{\infty} \lambda_k^{-1} \quad (4)$$

be divergent.

Proof. Consider the partial sum of the series (3)

$$S_n(t) = p_0(t) + p_1(t) + \dots + p_n(t) \quad (5)$$

From the birth equations it follows that

$$S'_n(t) = -\lambda_n p_n(t)$$

From this we find that

$$1 - S_n(t) = \lambda_n \int_0^t p_n(t) dt \quad (6)$$

(if in place of the initial condition $p_0(0)=1$ we take a different one, namely $p_i(0)=1$, then Equation (6) holds for $n \geq i$).

Since all terms of the sum (5) are nonnegative, for every fixed value of t the sum $S_n(t)$ does not diminish with increasing n . Consequently, the limit

$$\lim_{n \rightarrow \infty} (1 - S_n(t)) = \mu(t) \quad (7)$$

exists.

By virtue of (6) we conclude that

$$\lambda_n \int_0^t p_n(t) dt \geq \mu(t)$$

From this it is clear that

$$\int_0^t S_n(z) dz \geq \mu(t) \left(\frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_n} \right)$$

Since for any t and n the inequality $S_n(t) \leq 1$ holds, it follows that

$$t \geq \mu(t) \left(\frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_n} \right)$$

If the series (4) is divergent, it follows from the last inequality that $\mu(t)$ should be 0 for all t . From (7) it now follows that the divergence of the series (4) leads to (3).

From (6) it is clear that

$$\lambda_n \int_0^t p_n(t) dt \leq 1$$

and, hence,

$$\int_0^t S_n(t) dt \leq \frac{1}{\lambda_0} + \frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_n}$$

In the limit, as $n \rightarrow \infty$, we get

$$\int_0^t [1 - \mu(t)] dt \leq \sum_{n=0}^{\infty} \lambda_n^{-1}$$

If $\mu(t) = 0$ for all t , then the left-hand side of the inequality is equal to t and since t is arbitrary, the series on the right-hand side diverges. The theorem is proved.

In Sec. 51, we had $\lambda_n = \lambda$ for $n \geq 0$. Consequently, the series (4) diverges and for all t the equality $\sum_{n=0}^{\infty} p_n(t) = 1$ holds.

The sum $\sum_{n=0}^{\infty} p_n(t)$ may be interpreted as the probability that during time t there will occur only a finite number of changes of state of the system. Thus, the difference

$$1 - \sum_{n=0}^{\infty} p_n(t)$$

should be interpreted as the probability of an infinite number of changes of state of the system during time t . In radioactive decay this possibility implies an avalanche type of disintegration.

Example 1. Stand-by Relief Without Repair. Imagine a system consisting of one main server and n stand-by servers. During a time interval $(t, t+h)$ the main server can break down with a probability $\lambda h + o(h)$, and each of the stand-by servers (the so-called stand-by relief) can break down with a probability $\lambda' h + o(h)$. A server that has failed drops out of the system. The main server that has broken down is immediately replaced by one of the stand-by servers. The system as a whole breaks down as soon as all servers, both main and stand-by, fail. Find the probability that at time t there are k servers in the system that have broken down (event E_k).

This is a case of pure birth. Here $\lambda_k = \lambda + (n-k)\lambda'$ ($0 \leq k \leq n$), $\lambda_{n+k} = 0$ ($k \geq 1$). Simple calculations yield the equations

$$p_k(t) = \frac{\lambda_0 \lambda_1 \dots \lambda_{k-1}}{k! \lambda'^k} e^{-\lambda_k t} (1 - e^{-\lambda' t})^k \quad (0 \leq k \leq n)$$

and

$$p_{n+1}(t) = \frac{\lambda_0 \lambda_1 \dots \lambda_{n-1} \lambda}{n! \lambda'^n} \int_0^t e^{-\lambda z} (1 - e^{-\lambda' z})^n dz$$

In particular, if $\lambda' = 0$ (nonloaded stand-by in which the servers do not break down) the following equalities hold:

$$p_k(t) = \frac{\lambda^k t^k}{k!} e^{-\lambda t} \quad (0 \leq k \leq n), \quad p_{n+1}(t) = 1 - \sum_{k=0}^n \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

When $\lambda' = \lambda$ (loaded stand-by in which the stand-by servers are loaded just like the main server)

$$p_k(t) = C_{n+1}^k e^{-(n+1-k)\lambda t} (1 - e^{-\lambda t})^k$$

Denote by ξ_k the lifetime of the k th element in an operating period. For nonloaded stand-by relief the length of life of the system is

$$\xi_1 + \xi_2 + \dots + \xi_n$$

Since the average period of operation of one server is equal to

$$\int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda}$$

the mean operating period of the system, in the case of "cold" stand-by relief, is equal to $\frac{n+1}{\lambda}$; that is, it is proportional to the total number of servers in the system.

The mean operating time of a system without breakdown when the stand-by relief is loaded is calculated in the following manner: note the times of successive breakdowns t_1, t_2, \dots, t_{n+1} and introduce the notations $\tau_1 = t_1$, $\tau_2 = t_2 - t_1$, $\tau_3 = t_3 - t_2$, \dots , $\tau_{n+1} = t_{n+1} - t_n$. Since all servers are in operation during the first interval, the probability that during time t none will fail is equal to $e^{-\lambda(n+1)t}$; the probability that in the second interval not a single server will break down during time t is equal to $e^{-n\lambda t}$, and so forth; finally, the probability that during time t there will be no breakdowns in the $(n+1)$ st interval is equal to $e^{-\lambda t}$. The mean operating time of the system is

$$\sum_{k=1}^{n+1} M\tau_k = \frac{1}{\lambda} \left(1 + \frac{1}{2} + \dots + \frac{1}{n+1} \right)$$

If n is great, then

$$1 + \frac{1}{2} + \dots + \frac{1}{n} \sim \ln n + C$$

where C is Euler's constant.

We see that with increasing number of stand-by servers the mean time of no-breakdown operation of the system increases much faster in the case of nonloaded stand-by relief than in the case of loaded relief.

In the case of a pure birth process the system of Equations (1)-(2) was solved very simply by successive integration, since the differential equations had the form of recursion relations. The general equations of the birth and death process are of a different structure and the successive determination of the functions $p_k(t)$ is no longer possible. The conditions of the existence and uniqueness of solutions of this system have been thoroughly studied in the works of Feller, Reuter, McGregor and Karlin. It was found that the equation

$$\sum_{k=0}^{\infty} p_k(t) = 1$$

holds for all t if the series

$$\sum_{k=1}^{\infty} \prod_{i=1}^k \frac{v_i}{\lambda_i} \quad (8)$$

is divergent.

If, in addition, the series

$$\sum_{k=1}^{\infty} \prod_{i=1}^k \frac{\lambda_{i-1}}{v_i} \quad (9)$$

is convergent, then for all k the limits

$$p_k = \lim_{t \rightarrow \infty} p_k(t) \quad (10)$$

exist.

In particular, this condition holds in all cases when, for $k \geq k_0$ beginning with some k_0 onwards, the inequality

$$\frac{\lambda_k}{v_{k+1}} \leq \alpha < 1$$

holds. Intuitively, these conditions are clear: what they mean is that the arrival of calls (demands) in the service system must not exceed the speed of service.

To determine the limits (10), it is necessary to solve a system of algebraic equations that is obtained from the system (1)-(2) if we put $p'_k(t) = 0$ and if we put p_k in place of $p_k(t)$. This system then has the form

$$\begin{aligned} -\lambda_0 p_0 + v_1 p_1 &= 0, \\ -(\lambda_k + v_k) p_k + \lambda_{k-1} p_{k-1} + v_{k+1} p_{k+1} &= 0 \quad (k \geq 1) \end{aligned} \quad (11)$$

We introduce the notation

$$z_k = -\lambda_k p_k + v_{k+1} p_{k+1}, \quad k = 0, 1, 2, \dots$$

With this notation, the equations take the form

$$z_0 = 0, \quad z_{k-1} - z_k = 0 \quad (\text{for } k \geq 1)$$

Whence it follows that for all k

$$z_k = 0$$

Consequently

$$p_k = \frac{\lambda_{k-1}}{v_k} p_{k-1} = \prod_{i=1}^k \frac{\lambda_{i-1}}{v_i} p_0 \quad (12)$$

The constant p_0 is determined from the normalization condition

$$\left(\sum_{k=0}^{\infty} p_k = 1 \right):$$

$$p_0 = \left[1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{\lambda_{i-1}}{\nu_i} \right]^{-1} \quad (13)$$

It is obvious that these formulas contain the earlier obtained Erlang formulas.

To illustrate the theory we consider the following examples.

Example 2. A Service System with a Waiting Line (Queue). A Poisson flow of demands with parameter (intensity) λ comes into n identical servers. A demand that arrives at a server requires a random service time with probability distribution $H(x) = 1 - e^{-\nu x}$. If at the time of arrival of a demand there is at least one free server, it serves immediately. If all servers are busy, the new arrivals form a waiting line. If there is a waiting line, then as soon as a server completes a service session it immediately switches to servicing the next demand in line. The problem is to find the probability of one or another number of demands being in the service system.

Our conditions are those developed in the theory in the present section; for our problem $\lambda_k = \lambda$ for all k , $\nu_k = k\nu$ for $k \leq n$ and $\nu_k = n\nu$ for $k \geq n$.

According to formulas (12) and (13) the stationary solutions for our problem are of the form:

$$p_k = \frac{\rho^k}{k!} p_0$$

for $k \leq n$ and

$$p_k = \frac{\rho^k}{n! n^{k-n}} p_0$$

for $k \geq n$; here $\rho = \frac{\lambda}{\nu}$ and

$$p_0 = \left[1 + \sum_{k=1}^n \frac{\rho^k}{k!} + \frac{\rho^{n+1}}{n! (n - \rho)} \right]^{-1}$$

for $\rho < n$.

It turns out that for $\rho \geq n$, $p_0 = 0$, and also at the same time $p_k = 0$ for all k . This result is very important and should be borne in mind in practical situations. In words it may be formulated as follows: *in all cases in which $\rho \geq n$, the length of the queue increases without bound with time.*

Example 3. Maintenance of Machines by a Team of Repairmen. A team of r repairmen services n machines of the same type ($r \leq n$). Each one of the machines may demand the attention of a repairman at random moments. The machines go out of commission independently of one another. The probability of dropping out of operation during the time interval $(t, t+h)$ is equal to $\lambda h + o(h)$. The probability that during time $(t, t+h)$ a machine will be put into operation again is equal to $\nu h + o(h)$. Each repairman can repair only one machine at a time; each machine is handled by only one repairman. The parameters λ and ν are independent of t and n and also of the number of machines undergoing repair. Find the probability that in a steady-state process of service there will be a certain number of machines idle at a given time.

By E_k we denote the event that k machines are out of commission at a given instant. It is obvious that our system can only be in the states E_0, E_1, \dots, E_n . It is easy to see that we have to do with a birth and death process for which $\lambda_k = (n-k)\lambda$ for $0 \leq k < n$, $\lambda_k = 0$ for $k \geq n$; $\nu_k = k\nu$ for $1 \leq k \leq r$, $\nu_k = r\nu$ for $k \geq r$. Formulas (12) and (13) yield the equations: for $1 \leq k \leq r$ ($\rho = \frac{\lambda}{\nu}$)

$$p_k = \frac{n!}{k! (n-k)!} \rho^k p_0$$

for $r \leq k \leq n$

$$p_k = \frac{n!}{r^{n-k} r! (n-k)!} \rho^k p_0$$

and

$$p_0 = \left[\sum_{k=0}^r \frac{n!}{k! (n-k)!} \rho^k + \sum_{k=r+1}^n \frac{n!}{r! r^{n-k} (n-k)!} \rho^k \right]^{-1}$$

In particular, for $r=1$

$$p_k = \frac{n!}{(n-k)!} \rho^k p_0,$$

$$p_0 = \left[\sum_{k=0}^n \frac{n!}{(n-k)!} \rho^k \right]^{-1}$$

A simple numerical calculation will serve to illustrate these formulas. Suppose eight machines are serviced by two repairmen. What is the best way to organize the work: put both repairmen in charge of all the machines so that the workman that is free at the moment attends any machine that has just stopped, or assign four definite machines to each repairman? The calculations are carried out on the assumption that $\rho=0.2$. The results are tabulated.

$n = 8, \quad r = 2$

Number of nonoperating machines	Number of machines awaiting servicing	Number of repairmen idle	p_k
0	0	2	0.2048
1	0	1	0.3277
2	0	0	0.2294
3	1	0	0.1417
4	2	0	0.0687
5	3	0	0.0275
6	4	0	0.0083
7	5	0	0.0017
8	6	0	0.0002

The number of machines idle at a given time for the reason that the repairmen are busy with other machines is

$$\sum_{k=2}^{\infty} (k-2) p_k = 0.3045$$

The total idling time of the machines (repair and waiting for repair) is equal to

$$\sum_{k=2}^{\infty} k p_k = 1.6875$$

The mean duration of free time of the repairmen is

$$2 \times 0.2048 + 1 \times 0.3277 = 0.7373$$

In other words, each repairman is idle during 0.3686 working day.

$n = 4, \quad r = 1$

Number of nonoperating machines	Number of machines awaiting servicing	Number of repairmen idle	p_k
0	0	1	0.3984
1	0	0	0.3189
2	1	0	0.1914
3	2	0	0.0760
4	3	0	0.0153

The mean time of unproductive idling of machines (waiting for onset of repair work) is

$$1 \times 0.1914 + 2 \times 0.0760 + 3 \times 0.0153 = 0.3893$$

The whole group of eight machines will lose 0.7886 working day, which means the loss of working time of the machines due to waiting for repairs will more than double compared with the first type of labour organization. The total loss of time for the four machines (waiting+repairs) is

$$1 \times 0.3189 + 2 \times 0.1914 + 3 \times 0.0760 + 4 \times 0.0153 = 0.9909$$

Thus all eight machines will lose 1.9818 working days. On the average, a repairman is free during 0.3984 working day which means he is less engaged, though the machines stand idle for a longer time.

Sec. 62. Single-Server Queueing System

For the case when a service system has only one server, the queueing problem may be solved on the basis of much broader premises than those lying at the root of birth and death processes. This case is of particular interest from the standpoint of applications since one often has to do precisely with one server, or the flow of demands is divided in advance between servers according to a definite principle that is independent of the load on the servers at the time of arrival of a demand. Following the terminology of telephony, the case of one server goes by the name of a *single-server (one-channel) system*.

Suppose that the arrival times of the demands $t_1, t_2, \dots, t_n, \dots$ are such that the quantities $z_0 = t_1, z_1 = t_2 - t_1, \dots, z_n = t_{n+1} - t_n, \dots$ are mutually independent and identically distributed. We take it that

$$F(t) = P\{z_r < t\}$$

and

$$a = Mz_r = \int_0^{\infty} t dF(t) < +\infty$$

Demands arriving in the system are serviced immediately if a server is free, and get in a waiting line if it is busy servicing an earlier arrival. Servicing demands occupies a random time γ_r , where r is the serial number of an arriving demand. We assume that

$$G(x) = P\{\gamma_r < x\}$$

and

$$b = M\gamma_r = \int_0^{\infty} x dG(x) < +\infty$$

Our problem is to find the distribution of waiting time before servicing has begun. We denote this quantity by w_r for the r th arrival. For the sake of definiteness let $w_1 = 0$; that is, the server is assumed to be free at the start.

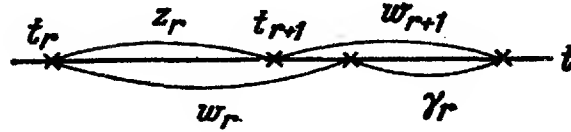


Fig. 21

It will readily be seen that we have the following equations:

$$w_{r+1} = \begin{cases} w_r + \gamma_r - z_r, & \text{if } w_r + \gamma_r - z_r > 0 \\ 0 & \text{if } w_r + \gamma_r - z_r \leq 0 \end{cases} \quad (1)$$

The first equation is well illustrated graphically in Fig. 21.

We put $u_r = \gamma_r - z_r$ and introduce the notation

$$U_r(x) = \mathbf{P} \{ \gamma_r - z_r < x \}$$

Since for $r \geq 1$, γ_r and z_r have distributions that are independent of r , the distribution of u_r is likewise independent of r for $r \geq 1$. We denote it by $U(x)$.

Further, denote the distribution function of w_r by $L_r(x)$. By virtue of the nonnegativity of the variables w_r for all r and $x \leq 0$, we have the equation

$$L_r(x) = 0$$

Since $w_1 = 0$ by assumption, for $x > 0$ we have

$$L_1(x) = 1$$

Relation (1) permits establishing the connection that exists between $L_r(x)$ and $L_{r+1}(x)$. Indeed, by (1), for $x > 0$, we have

$$\begin{aligned} L_{r+1}(x) &= \mathbf{P} \{ w_{r+1} < x \} = \mathbf{P} \{ w_{r+1} = 0 \} + \mathbf{P} \{ 0 < w_{r+1} < x \} = \\ &= \mathbf{P} \{ w_r + u_r \leq 0 \} + \mathbf{P} \{ 0 < w_r + u_r < x \} = \\ &= \mathbf{P} \{ w_r + u_r < x \} \end{aligned} \quad (2)$$

The variables w_r and u_r are independent, and so

$$L_{r+1}(x) = \int_{-\infty}^{\infty} L_r(x-v) dU_r(v) = \int_{-\infty}^x L_r(x-v) dU_r(v) \quad (3)$$

From this it is possible, knowing $L_1(x)$ and $U(x)$, to determine successively the functions $L_2(x), L_3(x), \dots$. In particular, when $x > 0$,

$$L_2(x) = \int_{-\infty}^x L_2(x-v) dU_1(v) = U(x)$$

and

$$L_3(x) = \int_{-\infty}^x L_2(x-v) dU_2(x) = \int_{-\infty}^x U(x-v) dU(v)$$

The last two equations permit formulating the following result: *the distribution of waiting time does not depend on the distributions $F(x)$ and $G(x)$ themselves, but only on the distribution $U(x)$.*

Note that the distribution $L_2(x)$ does not coincide fully with the distribution $U(x)$ since $L_2(x) = 0$ for $x \leq 0$, whereas it is possible to find $x < 0$ such that for them $U(x) > 0$. The probability that the second (in order of arrival) demand will be serviced without waiting is equal to $L_2(+0) - L_2(-0) = U(+0)$. The relation (2) yields more; namely, since $w_r > 0$ for any r , we can write the following sequence of equations for $x > 0$:

$$\begin{aligned} L_2(x) &= P\{u_1 < x\} \\ L_3(x) &= P\{u_1 + u_2 < x, u_2 < x\} \\ L_4(x) &= P\{u_1 + u_2 + u_3 < x, u_2 + u_3 < x, u_3 < x\} \\ &\dots \dots \dots \\ L_{r+1}(x) &= P\left\{\sum_{j=1}^r u_j < x, s=1, 2, \dots, r\right\} = \\ &= P\left\{\sum_{j=1}^s u_j < x, s=1, 2, \dots, r\right\} \end{aligned}$$

Let us now consider the behaviour of $L_r(x)$ as $r \rightarrow \infty$. It will readily be seen that for any x , as $r \rightarrow \infty$, the functions $L_r(x)$ tend to a limiting value. Indeed, consider for this purpose the event E_r that

$$\sum_{j=1}^s u_j < x \text{ for all } s \leq r$$

It is obvious that every E_r implies events E_s with smaller subscripts. If by E we denote the event

$$\sum_{j=1}^s u_j < x \text{ for all } s \geq 1$$

then, by virtue of the continuity axiom,

$$\lim_{r \rightarrow \infty} L_{r+1}(x) = \lim_{r \rightarrow \infty} \mathbf{P} \{E_r\} = \mathbf{P} \{E\}$$

If we introduce the notation

$$L(x) = \mathbf{P} \{E\} = \mathbf{P} \left\{ \sum_{j=1}^s u_j < x, s \geq 1 \right\}$$

then according to the preceding discussion Equation (3) will pass into

$$L(x) = \int_{-\infty}^x L(x-z) dU(z) \quad (4)$$

which, by an obvious change of variables, becomes

$$L(x) = \int_0^{\infty} L(y) dU(x-y) \quad (5)$$

Now let us investigate the behaviour of the function under various assumptions relative to the function $U(x)$, more precisely, relative to its first moment (the expectation of the variable $u_i = \gamma_i - z_i$). We know that according to the law of large numbers we have the equation

$$\mathbf{P} \left\{ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n u_i = \mathbf{M}u_1 = b - a \right\} = 1 \quad (6)$$

From this, for $\mathbf{M}u_1 > 0$, we conclude that with probability one there is an n_0 such that for all $n > n_0$ the inequality

$$\frac{1}{n} \sum_{i=1}^n u_i > \frac{1}{2} \mathbf{M}u_1$$

holds.

For every $x > 0$ there will be an n so large that $\frac{n}{2} \mathbf{M}u_1 > x$. The preceding inequality shows that with probability one we have the inequality $\sum_{i=1}^n u_i > x$ for any x and all sufficiently large n . This means that for any $x > 0$

$$L(x) = 0$$

The actual meaning of this equality is: with probability one the waiting time of the n th demand that has arrived in the service

system exceeds any $x > 0$ as n increases to infinity (in other words, as the time of operation of the service system approaches infinity).

Now let $Mu_1 < 0$, that is, $b < a$; on the average, the service time is less than the interval between successive arrivals of demands for service. We will prove that in this case $L(x)$ is a distribution function, that is, that $L(+\infty) = 1$. Indeed, by virtue of (6), an n_0 will be found such that

$$P \left\{ \sum_{i=1}^n u_i < 0 \text{ when } n > n_0 \right\} > 1 - \frac{\delta}{2}$$

But for every δ an x may be found such that

$$P \left\{ \sum_{i=1}^n u_i < x \text{ when } 1 \leq n \leq n_0 \right\} > 1 - \frac{\delta}{2}$$

Thus, for every $\delta > 0$ there will be an x such that

$$P \left\{ \sum_{i=1}^n u_i < x \text{ when } n \geq 1 \right\} > 1 - \delta$$

We therefore conclude that

$$L(+\infty) = \lim_{x \rightarrow \infty} P \left\{ \sum_{i=1}^n u_i < x; n \geq 1 \right\} = 1$$

The case of $Mu_i = 0$ requires more profound methods of investigation than the strong law of large numbers. It turns out that here, with the exception of the possibility of $u_i = 0$, when the service time is exactly equal to the interval prior to the arrival of the next demand, for every $x > 0$ we have the equality $L(x) = 0$. In other words, if on the average the service time is equal to the mean duration of the time interval between two successive arrivals of demands for service, then, with the exception of the trivial case mentioned above, the waiting line for service will increase without bound with time. This result is important both for theory and applications.

Now let us solve Equation (5). To do this, we compute the characteristic function of the integral

$$S(x) = \int_{-\infty}^x L(x-y) dU(y)$$

in two different ways. In the process, we introduce an additional assumption: for $x > 0$

$$F(x) = 1 - e^{-\lambda x}$$

in other words, we assume the traffic of demands for service to be elementary with parameter λ .

Since $S(x)$ is the distribution of the algebraic sum $w - z_1 + \gamma$, the characteristic function of $S(x)$ is

$$s(t) = \frac{\lambda}{\lambda + it} g(t) l(t)$$

Here, $g(t)$ and $l(t)$ denote, respectively, the characteristic functions for the distributions $G(x)$ and $L(x)$, and $\frac{\lambda}{\lambda + it}$, as will readily be seen, is the characteristic function of the variable $-z$.

By definition,

$$U(x) = \mathbf{P}\{\gamma - z < x\} = \lambda \int_0^{\infty} G(x+y) e^{-\lambda y} dy$$

According to (5),

$$S(x) = L(x)$$

for $x > 0$. When $x \leq 0$

$$S(x) = \int_{-\infty}^0 L(-z) d_z U(x+z) = \lambda \int_{-\infty}^0 L(-z) d \int_0^{\infty} G(x+y+z) e^{-\lambda y} dy$$

But x and y are negative, and so $G(x+y+z) = 0$ in the interval y from 0 to $-(x+z)$. Thus,

$$\begin{aligned} S(x) &= \lambda \int_{-\infty}^0 L(-z) d \int_{-(x+z)}^{\infty} G(x+y+z) e^{-\lambda y} dy = \\ &= \lambda \int_{-\infty}^0 L(-z) d \int_0^{\infty} G(v) e^{-\lambda(v-z-x)} dy = ce^{\lambda x} \end{aligned}$$

where

$$c = \int_{-\infty}^0 L(-z) \lambda^2 e^{\lambda z} dz \int_0^{\infty} G(v) e^{-\lambda v} dv$$

We now find

$$\begin{aligned} s(t) &= \int_{-\infty}^{\infty} e^{itx} dS(x) = c\lambda \int_{-\infty}^0 e^{itx + \lambda x} dx + \int_0^{\infty} e^{itx} dL(x) - c = \\ &= \frac{c\lambda}{\lambda + it} + l(t) - c \end{aligned}$$

Equating the two expressions for $s(t)$, we find that

$$l(t) = \frac{cti}{it + \lambda(1 - g(t))} = \frac{c}{1 + \lambda \frac{1 - g(t)}{it}}$$

Now compute the constant c . To do this, note that

$$l(0) = 1 = \lim_{t \rightarrow 0} \frac{c}{1 + \lambda \frac{1-g(t)}{it}} = \frac{c}{1 + \lambda \frac{g'(0)}{i}} = \frac{c}{1-b\lambda}$$

And, finally,

$$l(t) = \frac{1-\lambda b}{1 + \lambda \frac{1-g(t)}{it}}$$

Differentiation of this formula leads to the formulas of expectation and variance of the variable w , which is the waiting time for service of an arrival (demand):

$$Mw = \frac{\lambda \mu_2}{2(1-\lambda b)} \left(\mu_2 = \int_0^\infty x^2 dG(x) \right)$$

and

$$Dw = (Mw)^2 + \frac{\lambda \mu_3}{3(1-\lambda b)} \left(\mu_3 = \int_0^\infty x^3 dG(x) \right)$$

This formula shows that $\sqrt{Dw} > Mw$, or that considerable fluctuations are possible in the waiting time.

The foregoing formulas were obtained by A. Ya. Khinchin; the theory developed in the first half of this section is due to D. Lindley.

Sec. 63. Limit Theorem for Flows

We have already stated that the overwhelming majority of studies in the theory of queues and in reliability theory proceed, at the present time, from the assumption that the incoming flow of demands (or, in reliability theory, the flow of breakdowns) is elementary. In a number of practically important cases the initial assumptions that in Sec. 51 served as a basis for deriving the form of an elementary flow do not follow from any consideration of the physical picture of the phenomenon. Indeed, in certain problems we find significant deviations of actual traffic from the elementary type. It would seem that such deviations, due to the enormous diversity of conditions under which actual phenomena occur, should be the rule and not the exception. However, it appears that great deviations are incomparably more rare than might be expected on the basis of a priori reasoning. The problem thus arises of determining the causes by virtue of which elementary traffic is so often in good agreement with the course of real flows. In recent years this problem has been investigated in a large number of works. We confine ourselves here to only one model that leads to Poisson flows (these include elementary flows as well).

Suppose that the observed flow is the sum of a large number of independent flows of small intensity. Then, as will be demonstrated, the overall flow will, under very broad conditions, be almost Poisson. Problems very often involve total flows. The traffic of calls at a telephone exchange may be regarded as the sum of flows of individual subscribers. The incoming traffic of cargo ships at a port is the sum of the flows of the ships leaving from various other ports. The flow of breakdowns of an intricate device is the sum of the flows of breakdowns of its elements. The flow of calls for emergency medical aid is also made up of a very large number of calls from individuals. The list of concrete examples could be extended, but there is no need since anyone could add large numbers of cases from familiar spheres of activity. The thing to note here is that of particular interest is a consideration of such total flows, the component flows of which are uniformly small in some definite sense. We shall now prove a limit theorem for total flows on the basis of the ideas and methods given in Chapter 9.

A stochastic process $X(t)$ will be called a *step process* if for any x and s ($0 < s < t$) the increments $X(t) - X(s)$ can take on only nonnegative integral values. We assume that $X(0) = 0$. This means that the process began only at time $t = 0$. The value of the process $X(t)$ may be interpreted to mean the number of occurrences of certain events during the time interval between 0 and t . Such events are telephone calls arriving at a telephone exchange, arrival of clients at a barber shop, breakdown of electronic equipment, and so forth. Note that step processes can only change at specific points at once, a whole number of units. This may be regarded as the simultaneous arrival of several demands for service. Actual situations of this kind are rather frequent: the arrival in a port of a number of barges towed by a single tugboat, the arrival at a hospital of an ambulance with several people injured in an automobile accident, and so on.

Let

$$X_n(t) = \sum_{r=1}^{k_n} X_{nr}(t)$$

where $X_{nr}(t)$ are mutually independent step processes. It is obvious that the process $X_n(t)$ is also a step process.

We will say that the sequence of processes $X_n(t)$ *converges weakly* to the process $X(t)$ if the distribution function of the vectors

$$X_n(t_1), X_n(t_2), \dots, X_n(t_k)$$

for any choice of n, t_1, t_2, \dots, t_k at each point of continuity converges to the value of the distribution function of the vector

$$X(t_1), X(t_2), \dots, X(t_k)$$

We say that the process $X(t)$ is a *Poisson process with leading function* $\Lambda(t)$ if it: (1) has independent increments in nonoverlapping intervals and (2) for all $s < t$ and for any nonnegative integral k ,

$$\mathbf{P}\{X(t) - X(s) = k\} = \frac{[\Lambda(t) - \Lambda(s)]^k}{k!} e^{-[\Lambda(t) - \Lambda(s)]}$$

The leading function $\Lambda(t)$ is nonnegative, continuous on the left, finite for every t , and for $t \leq 0$ satisfies the equality $\Lambda(t) = 0$. For the elementary flow studied in Sec. 51, $\Lambda(t) = \lambda t$.

Let us introduce the following notation:

$$p_{nr}(k; s, t) = \mathbf{P}\{X_{nr}(t) - X_{nr}(s) = k\}, \quad k = 0, 1, 2, \dots \quad (1)$$

$$\Lambda_n(s, t) = \sum_{r=1}^{k_n} p_{nr}(1; s, t) \quad (2)$$

$$B_n(s, t) = \sum_{r=1}^{k_n} [1 - p_{nr}(0; s, t) - p_{nr}(1; s, t)] \quad (3)$$

Of the processes $X_{nr}(t)$ ($1 \leq r \leq k_n$) we will say that they are *infinitesimal* if for every fixed t

$$\lim_{n \rightarrow \infty} \max_{1 \leq r \leq k_n} [1 - p_{nr}(0; 0, t)] = 0 \quad (4)$$

In other words, the processes $X_{nr}(t)$ are infinitesimal if for any $\varepsilon > 0$ and an arbitrary fixed t it is possible to indicate an n such that for all r at once

$$\mathbf{P}\{X_{nr}(t) > \varepsilon\} < \varepsilon$$

We will now state and prove a limit theorem under conditions due to B. I. Grigelionis.

Theorem. *For the convergence of the sums*

$$X_n(t) = \sum_{r=1}^{k_n} X_{nr}(t)$$

of mutually independent infinitesimal processes $X_{nr}(t)$ to the Poisson process with leading function $\Lambda(t)$, it is necessary and sufficient that for any fixed s and t ($s < t$) the following relations should hold:

$$\lim_{n \rightarrow \infty} \Lambda_n(s, t) = \Lambda(t) - \Lambda(s) \quad (5)$$

and

$$\lim_{n \rightarrow \infty} B_n(0, t) = 0 \quad (6)$$

Proof. Proof of the necessity of the hypothesis of the theorem is based on the following proposition of the theory of summation

of independent random variables. If the independent random variables $x_{n1}, x_{n2}, \dots, x_{nk_n}$ are infinitesimal, that is, for any $\varepsilon > 0$ and as $n \rightarrow \infty$

$$\sup_{1 \leq k \leq k_n} \mathbf{P} \{ |x_{nk}| > \varepsilon \} \rightarrow 0$$

then in order that the distribution functions of the sums

$$s_n = x_{n1} + x_{n2} + \dots + x_{nk_n}$$

should converge to the Poisson distribution

$$\mathbf{P}(x) = \sum_{0 \leq k < x} \frac{\lambda^k}{k!} e^{-\lambda}$$

as n tends to infinity, it is necessary and sufficient that the following conditions be fulfilled: for every ε ($0 < \varepsilon < 1$) as n tends to infinity,

$$(1) \quad \sum_{k=1}^{k_n} \int_{R_\varepsilon} dF_{nk}(x) \rightarrow 0$$

$$(2) \quad \sum_{k=1}^{k_n} \int_{|x-1| < \varepsilon} dF_{nk}(x) \rightarrow \lambda$$

$$(3) \quad \sum_{k=1}^{k_n} \int_{|x| < \varepsilon} x dF_{nk}(x) \rightarrow 0$$

$$(4) \quad \sum_{k=1}^{k_n} \left[\int_{|x| < \varepsilon} x^2 dF_{nk}(x) - \left(\int_{|x| < \varepsilon} x dF_{nk}(x) \right)^2 \right] \rightarrow 0$$

Here, the following notations are introduced: $F_{nk}(x) = \mathbf{P} \{ X_{nk} < x \}$, R_ε is the region obtained from the real infinite line by eliminating intervals $|x| < \varepsilon$ and $|x-1| < \varepsilon$.

It will be noted that in the Grigelionis theorem we have to put

$$\begin{aligned} \lambda &= \Lambda(t) - \Lambda(s) \\ p_{nk}(1; s, t) &= \int_{|x-1| < \varepsilon} dF_{nk}(x) \\ 1 - p_{nk}(0; s, t) - p_{nk}(1; s, t) &= \int_{R_\varepsilon} dF_{nk}(x) \end{aligned}$$

It is now clear that the first and second conditions of the foregoing theorem on convergence to the Poisson distribution coincide exactly with the conditions (5) and (6). The third and fourth conditions of this theorem are automatically fulfilled for step processes, since in the interval $|x| < \varepsilon$ their distribution functions have only one point of increase $x=0$.

Thus, the necessity of the Grigelionis conditions follows from the fact that if the processes under consideration converge, then their one-dimensional distributions should converge as well.

We shall now prove that the conditions of the theorem are sufficient. To do this, it is obviously sufficient to demonstrate that the conditions (5) and (6) ensure both asymptotic independence of increments of the process $X_n(t)$ and convergence of one-dimensional distributions to the corresponding Poisson distributions. Incidentally, the second part of our program follows completely from the theorem we formulated concerning the convergence of the distribution functions of sums to the Poisson distribution.

Let us consider the vectors

$$\bar{l} = (l_1, l_2, \dots, l_m) \text{ where } l_v \geq 0 \text{ are integers}$$

$$\bar{0} = \underbrace{(0, 0, \dots, 0)}_m$$

$$\bar{l}_v = \underbrace{(0, 0, \dots, 0)}_{v-1}, 1, \underbrace{(0, 0, \dots, 0)}_{m-v}$$

$$\bar{T} = (t_0, t_1, \dots, t_m), \quad 0 \leq t_0 < t_1 < \dots < t_m \text{ are arbitrary real numbers}$$

$$\bar{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m)$$

$$\bar{X}_{nr}(\bar{T}) = (X_{nr}(t_1) - X_{nr}(t_0), \dots, X_{nr}(t_m) - X_{nr}(t_{m-1}))$$

$$\bar{X}_n(\bar{T}) = \sum_{r=1}^{k_n} \bar{X}_{nr}(\bar{T})$$

We also introduce the following supplementary notations:

$$\begin{aligned} &(\bar{\alpha}, \bar{\beta}) \sum_{i=1}^m \alpha_i \beta_i \\ &\rho_{nr}(\bar{l}, \bar{T}) = \mathbf{P} \{ \bar{X}_{nr}(\bar{T}) = \bar{l} \} \\ &f_{nr}(\bar{\alpha}, \bar{T}) = \mathbf{M} \exp i(\bar{\alpha}, \bar{X}_{nr}(\bar{T})) \\ &f_n(\bar{\alpha}, \bar{T}) = \mathbf{M} \exp i(\bar{\alpha}, \bar{X}_n(\bar{T})) \end{aligned}$$

For the distributions of the vectors $\bar{X}_{nr}(\bar{T})$ to converge to the corresponding distributions of the Poisson process as $n \rightarrow \infty$ it is sufficient that their characteristic functions converge. Let us try to detect this.

By virtue of the independence of the processes $\bar{X}_{nr}(\bar{T})$, we have

$$f_n(\bar{\alpha}, \bar{T}) = \prod_{k=1}^{k_n} f_{nr}(\bar{\alpha}, \bar{T})$$

But

$$f_{nr}(\bar{\alpha}, \bar{T}) = \sum_{\bar{l}} p_{nr}(\bar{l}, \bar{T}) e^{i(\bar{\alpha}, \bar{l})} = 1 + \sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) (e^{i(\bar{\alpha}, \bar{l})} - 1)$$

where the symbol $\sum_{\bar{l}}$ denotes summation over all possible integral vectors \bar{l} with nonnegative components.

For small x

$$1 + x = \exp [x + O(x^2)]$$

therefore,

$$\begin{aligned} f_{nr}(\bar{\alpha}, \bar{T}) &= \exp \left\{ \sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) (e^{i(\bar{\alpha}, \bar{l})} - 1) + O \left[\left(\sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) \right)^2 \right] \right\} = \\ &= \exp \left\{ \sum_{v=1}^m p_{nr}(\bar{l}_v, \bar{T}) (e^{i\alpha_v} - 1) + O \left(\sum_{\substack{\bar{l} \neq \bar{0}, \bar{l}_v \\ v=1, \dots, m}} p_{nr}(\bar{l}, \bar{T}) \right) + \right. \\ &\quad \left. + O \left[\left(\sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) \right)^2 \right] \right\} \end{aligned}$$

It is obvious that

$$\begin{aligned} \sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) &= 1 - \mathbf{P} \{ X_{nr}(t_m) - X_{nr}(t_0) = 0 \} \leq \\ &\leq 1 - \mathbf{P} \{ X_{nr}(t_m) = 0 \} = 1 - p_{nr}(0; 0, t_m) \\ \sum_{\substack{\bar{l} \neq \bar{0}, \bar{l}_v \\ v=1, \dots, m}} p_{nr}(\bar{l}, \bar{T}) &= \mathbf{P} \{ X_{nr}(t_m) - X_{nr}(t_0) \geq 2 \} \leq \\ &\leq \mathbf{P} \{ X_{nr}(t_m) \geq 2 \} \end{aligned} \quad (7)$$

$$\sum_{\bar{l} \neq \bar{0}} p_{nr}(\bar{l}, \bar{T}) \leq \sum_{v=1}^m p_{nr}(\bar{l}_v, \bar{T}) + \mathbf{P} \{ X_{nr}(t_m) \geq 2 \}$$

We note that

$$\begin{aligned} p_{nr}(1; t_{v-1}, t_v) - p_{nr}(\bar{l}_v, \bar{T}) &= \\ &= \mathbf{P} \{ X_{nr}(t_v) - X_{nr}(t_{v-1}) = 1, X_{nr}(t_{v-1}) - X_{nr}(t_0) + \\ &\quad + X_{nr}(t_m) - X_{nr}(t_v) \neq 0 \} \leq \mathbf{P} \{ X_{nr}(t_m) \geq 2 \} \end{aligned} \quad (8)$$

The relations (7) and (8) permit us to rewrite in different form the earlier found representation of the function $f_{nr}(\bar{\alpha}, \bar{T})$, namely

$$\begin{aligned} f_{nr}(\bar{\alpha}, \bar{T}) &= \exp \left\{ \sum_{v=1}^m p_{nr}(1; t_{v-1}, t_v) (e^{i\alpha_v} - 1) + \right. \\ &\quad \left. + O[\mathbf{P} \{ X_{nr}(t_m) \geq 2 \}] + O \left[(1 - p_{nr}(0; 0, t_m)) \sum_{v=1}^m p_{nr}(1; t_{v-1}, t_v) \right] \right\} \end{aligned}$$

From this we conclude that

$$f_n(\bar{\alpha}, \bar{T}) = \exp \left\{ \sum_{v=1}^m \Lambda_n(t_{v-1}, t_v) (e^{i\alpha_v} - 1) + \right. \\ \left. + O[B_n(0, t_m)] + O \left[\max_{1 \leq r \leq k_n} (1 - p_{nr}(0; 0, t_m)) \right] \right\}$$

The conditions of the theorem now lead to the following limit relationship: as n tends to infinity,

$$f_n(\bar{\alpha}, \bar{T}) \rightarrow \prod_{v=1}^m \exp \{ [\Lambda(t_v) - \Lambda(t_{v-1})] (e^{i\alpha_v} - 1) \}$$

thus proving the theorem.

The theorem that has just been proved permits us to obtain a large number of corollaries if we specialize the assumptions concerning the terms of the step processes. Prior to Grigelionis, A. Ya. Khinchin and G. A. Ososkov investigated the conditions of convergence of a total process to an elementary one on the assumption that the component processes are ordinary and their increments are stationary. The result which they obtained signifies qualitatively that if the component processes are independent, ordinary and stationary, then their infinitesimal character (given certain supplementary general conditions of a quantitative nature) practically ensures that such processes are asymptotically close to elementary processes. This result is of interest both to theory and to application.

Sec. 64. Elements of the Theory of Stand-by Systems

In present-day technology, reliability of equipment is increased by employing the method of *stand-by systems*, that is, the introduction of extra components, units and entire assemblies. The purpose of the supplementary devices is to take over operation if the basic systems break down. Depending on the state of the stand-by equipment, one distinguishes loaded, nonloaded and partially loaded relief. In the case of loaded relief, the stand-by unit is in the same state as the operating unit and for this reason has the same intensity of breakdowns. In the partially loaded case, the stand-by device is loaded, but not so fully as the main equipment and for this reason has a different breakdown intensity. A stand-by unit that is not loaded does not, naturally, suffer breakdown. The spare wheel of an automobile is a typical example of nonloaded relief. Quite naturally, loaded and nonloaded relief are special cases of partially loaded relief.

Numerous problems arise in the theory of stand-by systems that differ only terminologically from the problems of queueing theory. It is therefore natural, in a brief chapter devoted to the theory of queues, to examine certain problems in the theory of stand-by systems.

In order to illustrate the effectiveness of stand-by systems, let us consider a small numerical example. Suppose the probability that a certain device will operate without breakdown during a specified period of time is 0.9. Four such devices are to operate independently. The probability that all four devices will operate without failure is $0.9^4 = 0.6561$. Now suppose we have one device in a state of loaded relief. The probability of at least four devices operating without failure during the specified time is equal to $0.9^5 + 5 \times 0.9^4 \times 0.1 = 0.91854$. The presence of two reserve units increases the probability of non-breakdown operation of the system (that is, the probability of maintaining at least four devices in operation during the entire specified period of time) to 0.98415.

Many interesting results are given in a paper by A. D. Solov'yev devoted to stand-by systems without repair. The following is one such result.

It is possible to have an entire device in reserve; for instance, a generator at a power station or a diesel locomotive at a railway junction; it is also possible to have in reserve a component of a system or even a single element. The question arises as to what is preferable, to have large units or single elements in reserve.

Theorem. *If the switching of stand-by devices (units, elements, etc.) is flawless, then both in the case of loaded and nonloaded relief, an increase in the scale of the stand-by system reduces non-breakdown operation of the whole system.*

Proof. It will readily be seen that it is sufficient for us to confine ourselves to the case when only two parts of a device are united and there is one stand-by unit for each part. Figure 22a depicts schematically the stand-by system of an expanded system, and Fig. 22b illustrates stand-by relief of the individual parts. Let us call each of the

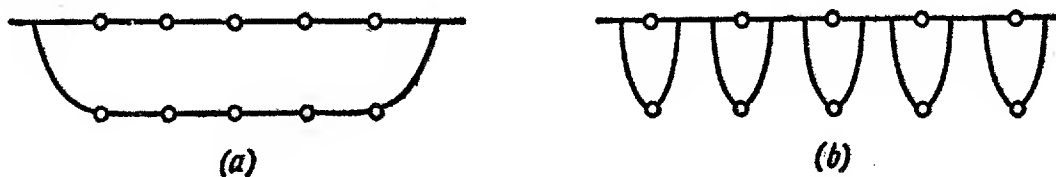


Fig. 22

five parts an element. By τ_1 and τ_2 we denote the duration of faultless operation of the basic elements and by τ'_1 and τ'_2 the duration of the corresponding time of the stand-by units. The distributions of these variables are arbitrary.

Let us denote by T_1 the duration of flawless operation of the supported system in the case of stand-by relief by a large unit and by T_2 the case of stand-by relief by single elements.

It is obvious that in the case of loaded relief we have the equations

$$T_1 = \max [\min (\tau_1, \tau_2), \min (\tau'_1, \tau'_2)]$$

and

$$T_2 = \min [\max (\tau_1, \tau'_1), \max (\tau_2, \tau'_2)]$$

Since

$$T_1 \leq \max (\tau_1, \tau'_1), T_1 \leq \max (\tau_2, \tau'_2)$$

it is clear that

$$T_1 \leq \min [\max (\tau_1, \tau'_1), \max (\tau_2, \tau'_2)] = T_2$$

For loaded relief, the assertion of the theorem has been proved.

For nonloaded relief we have

$$T_1 = \min (\tau_1, \tau_2) + \min (\tau'_1, \tau'_2)$$

and

$$T_2 = \min [\tau_1 + \tau'_1, \tau_2 + \tau'_2]$$

Since

$$T_1 \leq \tau_1 + \tau'_1, T_1 \leq \tau_2 + \tau'_2$$

we have the inequality

$$T_1 \leq \min [\tau_1 + \tau'_1, \tau_2 + \tau'_2] = T_2$$

This proves the theorem for nonloaded relief.

To increase the effectiveness of stand-by systems, devices that have failed are repaired. Let us investigate the effect of repair on increasing the reliability. We confine ourselves to the case of one basic and one reserve system.

Let us assume that the following conditions are fulfilled:

- (1) on breakdown of the basic device, the stand-by unit immediately takes up the load;
- (2) the device that has failed undergoes repair immediately;
- (3) the repairs fully restore the properties of the basic device that failed;
- (4) the repair time is a random variable with a distribution function $G(x)$;
- (5) the repaired device becomes a stand-by unit;
- (6) the period of faultless operation of the device is random and is distributed in accord with the law $F(x) = 1 - \exp(-\lambda x)$ ($\lambda > 0$) for the basic device and in accord with the law $F_1(x) = 1 - \exp(-\lambda_1 x)$ ($\lambda_1 \geq 0$) for the stand-by device. In particular, if the stand-by unit is loaded, then $\lambda_1 = \lambda$ and if it is nonloaded, then $\lambda_1 = 0$.

We shall say that our system (basic unit plus stand-by unit) breaks down if both devices go out of commission at the same time. Denote by $R(x)$ the probability that the system will operate flawlessly for a time greater than x . Let us also introduce the Laplace transforms

$$g(s) = \int_0^{\infty} e^{-sx} dG(x), \quad \varphi(s) = - \int_0^{\infty} e^{-sx} dR(x)$$

Theorem. Under the conditions (1) to (6), the function $R(x)$ satisfies the integral equation

$$R(x) = \exp [-(\lambda + \lambda_1)x] + (\lambda + \lambda_1) e^{-\lambda x} \int_0^x e^{-\lambda_1 z} [1 - G(x-z)] dz + \\ + (\lambda + \lambda_1) \int_0^x \int_0^{x-y} e^{-(\lambda + \lambda_1)y - \lambda z} R(x-y-z) dG(z) dy \quad (1)$$

In terms of Laplace transforms, the solution of this equation is given by the formula

$$\varphi(s) = \frac{\lambda(\lambda + \lambda_1)[1 - g(\lambda + s)]}{(\lambda + s)[s + (\lambda + \lambda_1)(1 - g(\lambda + s))]} \quad (2)$$

Proof. The event we are interested in—flawless operation of the system during time from 0 to x —is decomposable into three mutually independent events:

1. During the time $(0, x)$ neither the basic nor the stand-by element fails. The probability of this is equal to $e^{-(\lambda + \lambda_1)x}$.

2. The first breakdown occurs prior to time x . The remaining element operates flawlessly up to time x . Repair of the element which has failed is not completed prior to time x . The probability of this event is equal to

$$\int_0^x (\lambda + \lambda_1) e^{-(\lambda + \lambda_1)z} e^{-\lambda(x-z)} [1 - G(x-z)] dz$$

3. The first breakdown occurs prior to time x , the repair of this element is completed also prior to time x , during the repair period, the remaining element was functional. From the time of repair to time x , the system functioned normally. The probability of this event is equal to

$$\int \int_{y+z < x} (\lambda + \lambda_1) e^{-(\lambda + \lambda_1)y} e^{-\lambda z} R(x-y-z) dy dG(z)$$

Equating $R(x)$ to the sum of the three enumerated probabilities yields Equation (1).

Employment of the most elementary properties of Laplace transforms converts (1) into the equation

$$\varphi(s) = \frac{\lambda + \lambda_1}{s + (\lambda + \lambda_1)} \left[\frac{\lambda}{\lambda + s} - \frac{\lambda}{s + \lambda} g(\lambda + s) + \varphi(s) g(\lambda + s) \right]$$

from which (2) follows immediately.

It will be noted that by virtue of the properties of the exponential distribution, the result obtained is immediately extended to the case when there are n operating devices and one stand-by

unit. All devices have the same properties, that is, they have the same distribution functions for operating time and repairs. It is merely necessary to replace λ by $n\lambda$ in the formulas (1) and (2).

It is easy to calculate that the expectation of the time of flawless operation of the system is equal to

$$a = - \left[\frac{d\varphi(s)}{ds} \right]_{s=0} = \frac{\lambda + (\lambda + \lambda_1)(1 - g(\lambda))}{\lambda(\lambda + \lambda_1)(1 - g(\lambda))} \quad (3)$$

In particular, for a nonloaded stand-by system, we have

$$a_1 = \frac{2 - g(\lambda)}{\lambda(1 - g(\lambda))} \quad (3')$$

and for a loaded stand-by system

$$a_2 = \frac{3 - 2g(\lambda)}{2\lambda(1 - g(\lambda))} \quad (3'')$$

In the cases that are of most practical interest, the mean duration of repairs is considerably less than the mean time of flawless operation of the device. In order to invest with precise meaning the results that may be detected here, we prove the following limit theorems.

Suppose that the function $G(x)$ depends on a certain parameter ν and for any $\varepsilon > 0$, as $\nu \rightarrow \infty$,

$$1 - G_\nu(\varepsilon) \rightarrow 0 \quad (4)$$

It is easy to see that the following relation immediately follows from (3):

$$g_\nu(\lambda) \rightarrow 1 \quad (5)$$

as ν tends to infinity.

The converse is also true: if for any $s > 0$ and as ν tends to infinity we have the relation $g_\nu(s) \rightarrow 1$, then for any $x > 0$, as ν tends to infinity, $G_\nu(x)$ tends to one.

Theorem. *If condition (4) holds, then the flow of failures of a reduplicated system [conditions (1) to (6) are also assumed to hold] tends to the elementary case, given the choice of a proper unit of time.*

Proof. Put

$$\alpha_\nu = \left(1 + \frac{\lambda_1}{\lambda}\right)(1 - g_\nu(\lambda))$$

Then, by virtue of (2),

$$\varphi_\nu(\alpha_\nu s) = \frac{\lambda^2 \frac{1 - g_\nu(\lambda + \alpha_\nu s)}{1 - g_\nu(\lambda)}}{(\lambda + \alpha_\nu s) \left(s + \lambda \frac{1 - g_\nu(\lambda + \alpha_\nu s)}{1 - g_\nu(\lambda)} \right)} \quad (6)$$

Note that

$$\frac{1-g_v(\lambda+\alpha_v s)}{1-g_v(\lambda)} = 1 + \frac{g_v(\lambda)-g_v(\lambda+\alpha_v s)}{1-g_v(\lambda)}$$

and that for $s > 0$

$$0 \leq g_v(\lambda) - g_v(\lambda + \alpha_v s) = \int_0^\infty e^{-\lambda x} (1 - e^{-\alpha_v x s}) dG_v(x) \leq s \alpha_v \int_0^\infty x e^{-\lambda x} dG_v(x)$$

But by virtue of (4)

$$\begin{aligned} \int_0^\infty x e^{-\lambda x} dG_v(x) &= \int_0^\varepsilon x e^{-\lambda x} dG_v(x) + \int_\varepsilon^\infty x e^{-\lambda x} dG_v(x) \leq \\ &\leq \varepsilon + \max_x x e^{-\lambda x} \int_\varepsilon^\infty dG_v(x) \leq \varepsilon + \frac{\varepsilon}{\lambda e} = A\varepsilon \end{aligned}$$

Thus, in any finite interval s

$$\frac{1-g_v(\lambda+\alpha_v s)}{1-g_v(\lambda)} = 1 + o(1) \quad (7)$$

uniformly in s .

Now from (6) and (7) it follows that as v tends to infinity uniformly in s

$$\varphi_v(\alpha_v s) \rightarrow \frac{\lambda}{\lambda + s}$$

By the familiar theorems on Laplace transforms this means that the distribution of the random variable $\frac{\gamma_v}{\alpha_v}$, where γ_v denotes the length of the interval between two successive failures of the system on the condition that the repair-time function is $G_v(x)$, converges to the distribution $1 - e^{-\lambda x}$, as asserted by the theorem.

To estimate the effect of repair on the operational effectiveness of a system it is natural to consider the ratio of the mean operational time of a system with repair to that without repair. The former is calculated from the formula (3), the latter from the formula

$$a_0 = \frac{2\lambda + \lambda_1}{\lambda(\lambda + \lambda_1)}$$

The effectiveness of repairs is

$$e_v = \frac{\lambda + (\lambda + \lambda_1)(1 - g_v(\lambda))}{(2\lambda + \lambda_1)(1 - g_v(\lambda))} \quad (8)$$

Let us now determine what effect the choice of the function $G_v(x)$ has on the value of e_v . In this case, we shall naturally take all

$G_v(x)$ participating in the comparison to have the same expectation, which is assumed to be equal to $\frac{1}{v}$. For this purpose, consider the following distribution functions:

I.

$G_v(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{2} & \text{for } 0 < x \leq \frac{2}{v} \\ 1 & \text{for } x > \frac{2}{v} \end{cases}$

II.

$G_v(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - e^{-vx} & \text{for } x > 0 \end{cases}$

III.

$G_v(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{v}{2}x & \text{for } 0 < x \leq \frac{2}{v} \\ 1 & \text{for } x > \frac{2}{v} \end{cases}$

IV.

$G_v(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{1}{2}(3v)^3 \int_0^x z^2 e^{-3vz} dz & \text{for } x > 0 \end{cases}$

V.

$G_v(x) = \begin{cases} 0 & \text{for } x \leq \frac{1}{v} \\ 1 & \text{for } x > \frac{1}{v} \end{cases}$

We confine ourselves to the case of nonloaded stand-by relief and in Table 13 give all calculations dealing with the effectiveness of repair for the enumerated distributions.

TABLE 13

$G_v(x)$		e_v			
		$v/\lambda=1$	2	4	10
I	$1 + \frac{1 + e^{-2\lambda/v}}{2(2 - e^{-2\lambda/v})}$	1.66	2.08	3.04	6.02
II	$1 + \frac{v}{2\lambda}$	1.50	2.00	3.00	6.00
III	$1 + \frac{v(1 - e^{-2\lambda/v})}{2[2\lambda - v(1 - e^{-2\lambda/v})]}$	1.38	1.86	2.85	5.84
IV	$1 + \frac{(3v)^3}{2[(\lambda + 3v)^3 - (3v)^3]}$	1.36	1.85	2.84	5.84
V	$1 + \frac{e^{-\lambda/v}}{2(1 - e^{-\lambda/v})}$	1.29	1.77	2.76	5.75

The above table gives an amazingly small spread of the effectiveness of repair for such utterly different distributions of repair periods that we chose. The somewhat greater effectiveness for the first two distributions is due to the fact that they have an appreciable possibility of repair within short periods of time. The fact that the last distribution requires one and the same time for any repair reduces the effectiveness somewhat. The fact that the figures given in the table are so close to each other follows from the theorem we are about to prove.

Suppose that

$$m_1(v) = \int_0^{\infty} x dG_v(x) = \frac{1}{v}, \quad m_2(v) = \int_0^{\infty} x^2 dG_v(x) < +\infty$$

and, as v tends to infinity,

$$\frac{m_2(v)}{m_1(v)} \rightarrow 0 \quad (9)$$

Theorem. *If the condition (9) is satisfied in addition to the conditions (1) to (6), then for large values of v the mean time of flawless operation of a system with stand-by relief is asymptotically equal to the mean time of the system under the assumption that $G_v(x) = 1 - e^{-vx}$.*

Proof. Since for any $x > 0$

$$|e^{-x} - (1 - x)| \leq \frac{x^2}{2}$$

it follows that

$$\int_0^{\infty} |e^{-\lambda x} - 1 + \lambda x| dG_v(x) \leq \frac{\lambda^2 m_2(v)}{2}$$

It will be noted that

$$1 - g_v(\lambda) = \int_0^{\infty} \lambda x dG_v(x) - \int_0^{\infty} (e^{-\lambda x} - 1 + \lambda x) dG_v(x)$$

By virtue of (9)

$$1 - g_v(\lambda) = \lambda m_1(v) [1 + o(1)] = \frac{\lambda}{v} (1 + o(1)) = \frac{\lambda}{\lambda + v} (1 + o(1))$$

Substituting this estimate into (8), we find

$$e_v = \frac{2\lambda + \lambda_1 + v}{\lambda(\lambda + v)} (1 + o(1)) \quad (10)$$

A simple calculation shows that for the distribution $G_v(x) = 1 - e^{-vx}$

$$e_v = \frac{2\lambda + \lambda_1 + v}{\lambda(\lambda + v)}$$

A comparison of (10) and (11) proves the theorem.

Note that the condition (9) is automatically satisfied for all distributions with finite variance for which the following equation is valid:

$$G_v(x) = G_1(vx)$$

This relation holds for many distributions of practical importance, such as the Weibull distribution:

$$G(x) = 1 - e^{-\lambda x^\alpha} \quad (\lambda > 0, \alpha > 0)$$

the gamma distribution:

$$G'(x) = cx^\alpha e^{-\beta x} \quad (\alpha > -1, \beta > 0)$$

and others.

Appendix

TABLE A.1. Values of $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	0	1	2	3	4	5	6	7	8	9
0.0	0.3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0.1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0.2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0.3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0.4	3683	3668	3653	3637	3621	3605	3589	3572	3555	3538
0.5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0.6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0.7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0.8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0.9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1.0	0.2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1.1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1.2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1.3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1.4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1.5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1.6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1.7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1.8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1.9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2.0	0.0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2.1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2.2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2.3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2.4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2.5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2.6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2.7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2.8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2.9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3.0	0.0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3.1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3.2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3.3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3.4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3.5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3.6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3.7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3.8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3.9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

TABLE A.2. Values of $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

[illegible]

TABLE A.3. Values of $P_k(a) = \frac{a^k e^{-a}}{k!}$

<div><div><div><div></div><div><i>a</i></div><div></div></div><div><div><i>k</i></div><div></div></div></div></div>	0.1	0.2	0.3	0.4	0.5	0.6
0	0.904837	0.818731	0.740818	0.670320	0.606531	0.548812
1	0.090484	0.163746	0.222245	0.268128	0.303265	0.329287
2	0.004524	0.016375	0.033337	0.053626	0.075816	0.098786
3	0.000151	0.001091	0.003334	0.007150	0.012636	0.019757
4	0.000004	0.000055	0.000250	0.000715	0.001580	0.002964
5		0.000002	0.000015	0.000057	0.000158	0.000356
6			0.000001	0.000004	0.000013	0.000035
7					0.000001	0.000003
<div><div><div><div></div><div><i>a</i></div><div></div></div><div><div><i>k</i></div><div></div></div></div></div>	0.7	0.8	0.9	1.0	2.0	3.0
0	0.496585	0.449329	0.406570	0.367879	0.135335	0.049787
1	0.347610	0.359463	0.365913	0.367879	0.270671	0.149361
2	0.121663	0.143785	0.164661	0.183940	0.270671	0.224042
3	0.028388	0.038343	0.049398	0.061313	0.180447	0.224042
4	0.004968	0.007669	0.011115	0.015328	0.090224	0.168031
5	0.000695	0.001227	0.002001	0.003066	0.036089	0.100819
6	0.000081	0.000164	0.000300	0.000511	0.012030	0.050409
7	0.000008	0.000019	0.000039	0.000073	0.003437	0.021604
8		0.000002	0.000004	0.000009	0.000859	0.008101
9				0.000001	0.000191	0.002701
10					0.000038	0.000810
11					0.000007	0.000221
12					0.000001	0.000055
13						0.000013
14						0.000003
15						0.000001

TABLE A.3 (continued)

<div><div><i>a</i></div><div><i>k</i></div></div>	4.0	5.0	6.0	7.0	8.0	9.0
0	0.018316	0.006738	0.002479	0.000912	0.000335	0.000123
1	0.073263	0.033690	0.014873	0.006383	0.002684	0.001111
2	0.146525	0.084224	0.044618	0.022341	0.010735	0.004998
3	0.195367	0.140374	0.089235	0.052129	0.028626	0.014994
4	0.195367	0.175467	0.133853	0.091226	0.057252	0.033737
5	0.156293	0.175467	0.160623	0.127717	0.091604	0.060727
6	0.104194	0.146223	0.160623	0.149003	0.122138	0.091090
7	0.059540	0.104445	0.137677	0.149003	0.139587	0.117116
8	0.029770	0.065278	0.103258	0.130377	0.139587	0.131756
9	0.013231	0.036266	0.068838	0.101405	0.124077	0.131756
10	0.005292	0.018133	0.041303	0.070983	0.099262	0.118580
11	0.001925	0.008242	0.022529	0.045171	0.072190	0.097020
12	0.000642	0.003434	0.011262	0.026350	0.048127	0.072765
13	0.000197	0.001321	0.005199	0.014188	0.029616	0.050376
14	0.000056	0.000472	0.002228	0.007094	0.016924	0.032384
15	0.000015	0.000157	0.000891	0.003311	0.009026	0.019431
16	0.000004	0.000049	0.000334	0.001448	0.004513	0.010930
17	0.000001	0.000014	0.000118	0.000596	0.002124	0.005786
18		0.000004	0.000039	0.000232	0.000944	0.002893
19		0.000001	0.000012	0.000085	0.000397	0.001370
20			0.000004	0.000030	0.000159	0.000617
21			0.000001	0.000010	0.000061	0.000264
22				0.000003	0.000022	0.000108
23				0.000001	0.000008	0.000042
24					0.000003	0.000016
25					0.000001	0.000006
26						0.000002
27						0.000001

TABLE A.4. Values of $\sum_{m=0}^k \frac{a^m e^{-a}}{m!}$

<div><div><div><div></div><div><i>a</i></div><div></div></div><div><div><i>k</i></div><div></div></div></div></div>	0.1	0.2	0.3	0.4	0.5	0.6
0	0.904837	0.818731	0.740818	0.670320	0.606531	0.548812
1	0.995321	0.982477	0.963063	0.938448	0.909796	0.878099
2	0.999845	0.998852	0.996400	0.992074	0.985612	0.976885
3	0.999996	0.999943	0.999734	0.999224	0.998248	0.996642
4	1.000000	0.999998	0.999984	0.999939	0.999828	0.999606
5	1.000000	1.000000	0.999999	0.999996	0.999986	0.999962
6	1.000000	1.000000	1.000000	1.000000	0.999999	0.999997
7	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
<div><div><div><div></div><div><i>a</i></div><div></div></div><div><div><i>k</i></div><div></div></div></div></div>	0.7	0.8	0.9	1.0	2.0	3.0
0	0.496585	0.449329	0.406570	0.367879	0.135335	0.049787
1	0.844195	0.808792	0.772483	0.735759	0.406006	0.199148
2	0.965858	0.952577	0.937144	0.919699	0.676677	0.423190
3	0.994246	0.990920	0.986542	0.981012	0.857124	0.647232
4	0.999214	0.998589	0.997657	0.996340	0.947348	0.815263
5	0.999909	0.999816	0.999658	0.999406	0.983437	0.916082
6	0.999990	0.999980	0.999958	0.999917	0.995467	0.966491
7	0.999998	0.999999	0.999997	0.999990	0.998904	0.988095
8	1.000000	1.000000	1.000000	0.999999	0.999763	0.996196
9				1.000000	0.999954	0.998897
10					0.999992	0.999707
11					0.999999	0.999928
12					1.000000	0.999983
13						0.999996
14						0.999999
15						1.000000

TABLE A.4 (continued)

<div><div><div><div><div></div><div><i>a</i></div></div></div><div><div><i>k</i></div><div></div></div></div></div>	4.0	5.0	6.0	7.0	8.0	9.0
0	0.018316	0.006738	0.002479	0.000912	0.000335	0.000123
1	0.091579	0.040428	0.017352	0.007295	0.003019	0.001234
2	0.238105	0.124652	0.061970	0.029636	0.013754	0.006232
3	0.433472	0.265026	0.151205	0.081765	0.042380	0.021228
4	0.628839	0.440493	0.285058	0.172991	0.099632	0.054963
5	0.785132	0.615960	0.445681	0.300708	0.191236	0.115690
6	0.889326	0.762183	0.606304	0.449711	0.313374	0.206780
7	0.948866	0.866628	0.743981	0.598714	0.452961	0.323896
8	0.978636	0.931806	0.847239	0.729091	0.592548	0.455652
9	0.991867	0.968172	0.916077	0.830496	0.716625	0.587408
10	0.997159	0.986305	0.957380	0.901479	0.815887	0.705988
11	0.999084	0.994547	0.979909	0.946650	0.888077	0.803008
12	0.999726	0.997981	0.991173	0.973000	0.936204	0.875773
13	0.999923	0.999202	0.996372	0.987188	0.965820	0.926149
14	0.999979	0.999774	0.998600	0.994282	0.982744	0.958533
15	0.999994	0.999931	0.999491	0.997593	0.991770	0.977964
16	0.999998	0.999980	0.999825	0.999041	0.996283	0.988894
17	0.999999	0.999994	0.999943	0.999637	0.998407	0.994680
18	0.999999	0.999998	0.999982	0.999869	0.999351	0.997573
19	0.999999	0.999999	0.999994	0.999955	0.999748	0.998943
20	1.000000	0.999999	0.999998	0.999985	0.999907	0.999560
21		1.000000	0.999999	0.999995	0.999967	0.999824
22			0.999999	0.999998	0.999989	0.999932
23			1.000000	0.999999	0.999997	0.999974
24				0.999999	0.999999	0.999990
25				1.000000	0.999999	0.999996
26					1.000000	0.999998
27						0.999999
28						1.000000

Bibliography

(STARRED ITEMS ARE IN RUSSIAN)

POPULAR

1. Borel, E., *Le hasard*, Paris, 2nd ed., 1948.
2. Borel, E., *Probability and Certainty*, New York, 1963.
- 3.* Gnedenko, B. V., *How Mathematics Studies Random Phenomena*, Izd. Akad. Nauk Ukr. S.S.R., Kiev, 1947.
- 4.* Gnedenko, B. V. and Khinchin, A. Ya., *An Elementary Introduction to Probability Theory*, 6th ed., Izd. "Nauka", 1964.
- 5.* Yaglom, A. M. and Yaglom, I. M., *Probability and Information*, 2nd ed., Fizmatgiz, 1960.

TEXTBOOKS AND MONOGRAPHS

6. Bartlett, M. S., *An Introduction to Stochastic Processes*, Cambridge, 1955.
- 7.* Bernstein, S. N., *Probability Theory*, 4th ed., Gostekhizdat, 1946.
8. Blackwell, D. and Girshick, M. A., *Theory of Games and Statistical Decisions*, New York, 1954.
9. Blanc-Lapierre, A. et Fortet, R., *Théorie des fonctions aléatoires*, Paris, 1953.
10. Chandrasekar, S., *Stochastic Problems in Physics and Astronomy*, Rev. Modern Phys., Vol. 15, 1943.
11. Chung, Kai-Lai, *Markov Chains with Stationary Transition Probabilities*, Springer, Berlin, 1960.
12. Cramér, H., *Random Variables and Probability Distributions*, Cambridge University Press, 2nd ed., 1961.
13. Doob, J. L., *Stochastic Processes*, New York, 1953.
- 14.* Dunin-Barkovsky, I. V. and Smirnov, N. V., *The Theory of Probability and Mathematical Statistics* (general part), GTTI, Moscow, 1955.
- 15.* Dynkin, E. B., *Fundamentals in the Theory of Markov Processes*, Fizmatgiz, 1959.
- 16.* Dynkin, E. B., *Markov Processes*, Fizmatgiz, 1963.
- 17.* Einstein and Smoluchowski, *Collection of Articles on the Theory of Brownian Motion*, ONTI, 1936.
18. Fisz, M., *Rachunek prawdopodobieństwa i statystyka matematyczna*, Warszawa, 1958.

19. Frechet, M., *Recherches théoriques modernes. Traite du calcul des probabilités*, Paris, 1937, t. I, II.
- 20.* Gilenko, N. D., *Problems in Probability Theory*, Uchpedgiz, 1943.
- 21.* Glivenko, V. I., *A Course in the Theory of Probability*, GONTI, 1939.
- 22.* Glivenko, V. I., *The Stieltjes Integral*, ONTI, 1936.
- 23.* Gnedenko, B. V. and Kolmogorov, A. N., *Limit Distributions for Sums of Independent Random Variables*, Gostekhizdat, 1949.
24. Grenander, U., *Probabilities on Algebraic Structures*, New York, 1963.
25. Hannan, E. J., *Time Series Analysis*, Mathuen and Co., London, 1960.
26. Harris, T. E., *The Theory of Branching Processes*, Springer-Verlag, 1963.
- 27.* Ito, K., *Stochastic Processes*, IL, "Matematika", Issue 1, 1960; Issue 2, 1963.
- 28.* Khinchin, A. Ya., *Basic Laws of Probability Theory*, GTTI, 1932.
- 30.* Khinchin, A. Ya., *Limit Laws for Sums of Independent Random Variables*, ONTI, 1938.
- 31.* Khinchin, A. Ya., *Mathematical Foundations of Statistical Mechanics*, Gostekhizdat, 1943.
- 32.* Khinchin, A. Ya., *Mathematical Foundations of Quantum Statistics*, Gostekhizdat, 1951.
- 33.* Khinchin, A. Ya., *Mathematical Methods in the Study of Queues*, Trudy Mat. inst. imeni V. A. Steklova, Izd. Akad. Nauk S.S.S.R., 1955.
- 34.* Kolmogorov, A. N., *Basic Concepts of Probability Theory*, ONTI, 1936.
- 35.* Kubilius, I., *Probabilistic Methods in Number Theory*, Vilnius, 1959.
36. Laning, J. H. and Battin, R. H., *Random Processes in Automatic Control*, New York, 1956.
37. Lévy, P., *Théorie de l'addition des variables aléatoires*, Paris, 1937.
38. Lévy, P., *Processus stochastiques et mouvement brownien*, Paris, 1948.
- 39.* Linnik, Yu. V., *Decompositions of Probability Distributions*, Izd. LGU, 1960.
40. Loève, M., *Probability Theory*, Princeton, 3rd ed., 1963.
- 41.* Markov, A. A., *The Calculus of Probabilities*, 4th ed., GIZ, 1924.
- 42.* Meshalkin, L. D., *Collection of Problems in Probability Theory*, Izd. MGU, 1964.
43. Mises, R. von, *Wahrscheinlichkeitsrechnung*, 1931.
44. Mises, R. von, *Probability, Statistics, and Truth*, New York, 1939.
45. Onicescu, O., Mihoc, G., Jonescu Tulcea, C. T., *Calculul Probabilităților și aplicatii*, București, 1956.
46. Parzen, E., *Modern Probability Theory and Its Applications*, John Wiley and Sons, Inc., New York, 1960.
47. Rényi, A., *Wahrscheinlichkeitsrechnung*, Deutsche Verlag der Wissenschaften, 1962.
48. Richter, H., *Wahrscheinlichkeitsrechnung*, Springer-Verlag, 1956.
- 49.* Romanovsky, V. I., *Discrete Markov Chains*, Gostekhizdat, 1949.
50. Rosenblatt, M., *Random Processes*, Oxford University Press, N. Y., 1962.
- 51.* Rozanov, Yu. A., *Stationary Stochastic Processes*, Fizmatgiz, 1963.
52. Saaty, T., *Elements of Queueing Theory with Its Applications*, McGraw-Hill Book Company, New York, 1961.
- 53.* Sarymsakov, T. A., *Fundamentals in the Theory of Markov Processes*, Gostekhizdat, 1954.
- 54.* Sirazhdinov, S. Kh., *Limit Theorems for Homogeneous Markov Chains*, Izd. Akad. Nauk Uz. S.S.R., 1955.
- 55.* Skorokhod, A. V., *Studies in the Theory of Stochastic Processes*, Izd. Kiev. Universiteta, 1961.
- 56.* Skorokhod, A. V., *Stochastic Processes with Independent Increments*, Izd. "Nauka", 1964.
57. Todhunter, J., *A History of the Mathematical Theory of Probability*, 1865.
58. Tortrat, A., *Calcul des probabilités*, Masson et Cie, Paris, 1963.
- 59.* Ventstzel, E. S., *Probability Theory*, 3rd ed., Izd. "Nauka", 1964.

JOURNALS

Chapter 1

- 60.* Bernstein, S. N., "On the Axiomatic Substantiation of Probability Theory", *Reports of Kharkov Math. Society*, Vol. 15 (1917).
- 61.* Khinchin, A. Ya., "Mises' Theory of Probability and the Principles of Physical Statistics", *Uspekhi fizich. nauk*, Vol. IX, Issue 2 (1929).
- 62.* Khinchin, A. Ya., "The Method of Arbitrary Functions and the Struggle Against Idealism in Probability Theory", in Collection of articles entitled *Philosophical Problems of Modern Physics*, Izd. Akad. Nauk S. S. S. R., 1952.
- 63.* Khinchin, A. Ya., "The Frequency Theory of Richard von Mises and Modern Ideas in Probability Theory", *Voprosy filosofii*, Nos. 1 and 2 (1961).
- 64.* Kolmogorov, A. N., "The Role of Russian Science in the Development of Probability Theory", *Uchen. zap. MGU*, Issue 91 (1947).
- 65. *Théorie des probabilités. Exposés sur ses fondement et ses applications*, Paris, Gautier-Villars, 1952.

Chapter 2

- 66.* Bernstein, S. N., "Once again on the Question of the Accuracy of the Laplace Limit Formula", *Izv. Akad. Nauk S.S.S.R.*, Vol. 7 (1943).
- 67. Feller, W., "On the Normal Approximation to the Binomial Distribution", *Ann. Math. Stat.*, Vol. XVI (1945).
- 68. Khinchin, A. Ya., "Über einen neuen Grenzwertsatz der Wahrscheinlichkeitsrechnung", *Math. Annal.*, Vol. 101 (1929).
- 69.* Prokhorov, Yu. V., "Asymptotic Behaviour of the Binomial Distribution", *Uspekhi matem. nauk*, Vol. 8, Issue 3, pp. 136-142 (1953).
- 70.* Smirnov, N. V., "On the Probabilities of Large Deviations", *Mat. Sbornik*, Vol. 40, No. 4 (1933).

Chapter 3

- 71.* Dobrushin, R. L., "Limit Theorems for Markov Chains of Two States", *Izv. Akad. Nauk S.S.S.R.*, Ser. mat., 17, pp. 291-330 (1953).
- 72.* Dobrushin, R. L., "Central Limit Theorem for Nonhomogeneous Markov Chains", *Probability Theory and Its Applications*, Vol. 1, Issue 1, pp. 72-89, Issue 4, pp. 365-425 (1956).
- 73. Doeblin, W., "Exposé de la théorie des chaines simples constante de Markoff a un nombre fini d'états", *Rev. math. de l'Union Interbalkanique*, II, I (1938).
- 74.* Kolmogorov, A. N., "Markov Chains with Countable Number of Possible States", *Bull. MGU*, Vol I, Issue 3 (1937).
- 75.* Markov, A. A., "Investigation of a Remarkable Case of Dependent Trials", *Izv. Ros. Akad. Nauk*, Vol. 1 (1907). See also appropriate chapters in the books of Bernstein and Feller given in the list of textbooks and monographs; also in Romanovsky and Frechet (Vol. 2). The books of Doob and Sarymsakov contain extensive bibliographies on Markov chains.

Chapter 4

- 76. Cramér, H., "Über eine Eigenschaft der normalen Verteilungsfunktion", *Math. Zeitschr.*, Vol. 41 (1936).
- 77.* Raikov, D. A., "On the Decomposition of the Laws of Gauss and Poisson", *Izv. Akad. Nauk S.S.S.R.*, Ser. mat., pp. 91-124 (1938).
- 78.* Skitovich, V. P., "Linear Forms of Independent Random Variables and the Normal Distribution Law", *Izv. Akad. Nauk S.S.S.R.*, 18 (1952) (1954).

Chapter 6

- 79.* Bernstein, S. N., "On the Law of Large Numbers", *Soobshch. Khark. Mat. Obshchestva*, Vol. XVI (1918).
- 80.* Chebyshev, P. L., "On Mean Quantities", *Mat. Sbornik*, Vol. 2 (1867); Complete Works, Vol. 2, 1948.
81. Hájek, I. and Rényi, A., "Generalization of an Inequality of Kolmogorov", *Acta Math. Acad. Sc. Hungarica*, t. VI, fasc. 3-4, pp. 281-283 (1955).
- 82.* Kolmogorov, A. N., "Sur la loi fort des grands nombres", *C. R. Acad. Sci.*, Paris, 191, 910-912 (1930).
- 83.* Prokhorov, Yu. V., "On the Strong Law of Large Numbers", *Doklady Akad. Nauk S.S.S.R.*, Vol. 69, No. 5 (1949).
84. Slutsky, E. E., "Über stochastische Asymptoten und Grenzwerte", *Metron* 5 (1925); Selected Works, Izd. Akad. Nauk S.S.S.R., 1960.

Chapter 7

- 85.* Gnedenko, B. V., "On Characteristic Functions", *Bull. MGU*, Vol. 1. Issue 5 (1937).
- 86.* Khinchin, A. Ya., "On a Criterion for Characteristic Functions", *Bull. MGU*, Vol. 1, Issue 5 (1937).
- 87.* Krein, M. G., "On the Representation of Functions by Means of Fourier-Stieltjes Integrals", *Uchen. zap. Kuibyshevsk. ped. inst.*, Issue 7 (1943).
- 88.* Raikov, D. A., "On Positive Definite Functions", *Doklady Akad. Nauk S.S.S.R.*, Vol., XXVI, No. 9, pp. 857-862 (1940).

Chapter 8

- 89.* Bernstein, S. N., "Extending the Limit Theorem of Probability Theory to Sums of Dependent Variables", *Uspekhi mat. nauk*, Issue 10 (1944).
- 90.* Chebyshev, P. L., "On Two Theorems Concerning Probability", *Zap. Akad. Nauk* (1887); Complete Works, Vol. 2, 1948.
91. Esseen, C. G., "Fourier Analysis of Distribution Functions. A Mathematical Study of the Laplace-Gaussian Law", *Acta Math.*, Vol. 77 (1945).
92. Feller, W., "Über den Zentralengrenzwertsatz der Wahrscheinlichkeitsrechnung", *Math. Zeitschr.*, Vol. 40 (1935).
- 93.* Gnedenko, B. V., "Elements of the Theory of Distribution Functions of Random Vectors", *Uspekhi mat. nauk*, Issue 10 (1944).
- 94.* Gnedenko, B. V., "On the Local Limit Theorem of Probability Theory", *Uspekhi mat. nauk*, Vol. III, Issue 3 (1948).
- 95.* Gnedenko, B. V., "The Local Limit Theorem for Densities", *Doklady Akad. Nauk S.S.S.R.*, Vol. 95, No. 1 (1954).
96. Lindeberg, J. W., "Eine neue Herleitung des Exponentialgesetz in der Wahrscheinlichkeitsrechnung", *Math. Zeitschr.*, Vol. 15 (1922).
- 97.* Linnik, Yu. V., "On the Accuracy of the Approach of Sums of Independent Random Variables to the Gaussian Distribution", *Izv. Akad. Nauk S.S.S.R.*, Vol. 11 (1947).
- 98.* Lyapunov, A. M., "Sur une proposition de la théorie des probabilités", *Bull. Acad. Sc. Péter.*, 13, (1900).
- 99.* Lyapunov, A. M., "Nouvelle forme du théoreme sur la limite des probabilités", *ibid.* (1901).
- 100.* Prokhorov, Yu. V., "Local Theorem for Densities", *Doklady Akad. Nauk S.S.S.R.*, Vol. 83, No. 6 (1952).

Chapter 9

101. Bavli, G. M., "Über einige Verallgemeinerungen der Grenzwertsatz der Wahrscheinlichkeitsrechnung", *Mat. Sbornik*, Vol. I (43), No. 6 (1936).

- 102.* Gnedenko, B. V., "On a Characteristic Property of Infinitely Divisible Distribution Laws", *Bull. MGU*, Vol. I, Issue 5 (1937).
- 103.* Gnedenko, B. V. "Limit Laws for Sums of Independent Random Variables", *Uspekhi mat. nauk*, Issue 10 (1944).
- 104.* Khinchin, A. Ya., "A New Derivation of a Formula by P. Lévy", *Bull. MGU*, Vol. I, Issue 1 (1937).

Chapter 10

- 105. Cramér, H., "On Harmonic Analysis in Certain Continuous Functional Spaces", *Ark. Mat. Astr. Fys.*, 28B, No. 12 (1942).
- 106.* Dubrovsky, V. M., "Generalizing the Theory of Purely Discontinuous Stochastic Processes", *Doklady Akad. Nauk S.S.S.R.*, Vol. XIX (1938).
- 107.* Dubrovsky, V. M., "An Investigation of Purely Discontinuous Stochastic Processes by the Method of Integro-Differential Equations", *Izv. Akad. Nauk S.S.S.R.*, Vol. 8 (1944).
- 108. Feller, W., "On the Theory of Stochastic Processes", *Uspekhi mat. nauk*, Issue 5 (1938).
- 109. Karhunen, K., "Über lineare Methoden in der Wahrscheinlichkeitsrechnung", *Ann. Acad. Sci. Fennicae*, A, I, No. 37, Helsinki (1947).
- 110.* Khinchin, A. Ya., "Correlation Theory of Stationary Stochastic Processes", *Uspekhi mat. nauk*, Issue 5 (1938).
- 111.* Kolmogorov, A. N., "Simplified Proof of the Ergodic Theorem of Birkhoff-Khinchin", *Uspekhi mat. nauk*, Issue 5 (1938).
- 112.* Kolmogorov, A. N., "On Analytical Methods in Probability Theory", *Uspekhi mat. nauk*, Issue 5 (1938).
- 113.* Kolmogorov, A. N., "Interpolation and Extrapolation of Stationary Random Sequences", *Izv. Akad. Nauk S.S.S.R.* (1941).
- 114.* Kolmogorov, A. N., "A Statistical Theory of Vibrations with Continuous Spectrum", *Jubil. Sbornik Akad. Nauk S.S.S.R.*, Part I (1947).
- 115.* Kolmogorov, A. N. and Dmitriev, N. A., "Branching Stochastic Processes", *Doklady Akad. Nauk S.S.S.R.*, Vol. 56, No. 1 (1947).
- 116.* Kolmogorov, A. N. and Sevastyanov, B. A., "Computation of Final Probabilities for Branching Stochastic Processes", *Doklady Akad. Nauk S.S.S.R.*, Vol. 56, No. 8 (1947).
- 117. Loève, M., "Sur les fonctions aléatoires stationnaires de second ordre", *Rev. Sci.*, 83, No. 5 (1945).
- 118. Loève, M., "Fonctions aléatoires à décomposition orthogonale exponentielle", *Rev. Sci.*, 84, No. 3 (1946).
- 119. Maruyama, G., "The Harmonic Analysis of Stationary Stochastic Processes", *Mem. Fac. Sc. Kyusyu Univ.*, A, 4, No. 1 (1949).
- 120.* Rozanov, Yu. A., "The Spectral Theory of Multidimensional Stationary Processes with Discrete Time", *Uspekhi mat. nauk*, Issue 2 (1958).
- 121.* Sevastyanov, B. A., "The Theory of Branching Stochastic Processes", *Uspekhi mat. nauk*, Vol. 6, Issue 6 (1951).
- 122.* Yaglom, A. M., "On the Question of Linear Interpolation of Stationary Stochastic Sequences and Processes", *Uspekhi mat. nauk*, Vol. IV, Issue 4 (1949).
- 123.* Yaglom, A. M., "Introduction to the Theory of Stationary Stochastic Functions", *Uspekhi mat. nauk*, Vol. VII, Issue 5 (1952).

Chapter 11

- 124.* Belyaev, Yu. K., "Line Markov Processes and Their Application to Problems in Reliability Theory", Transactions of the 6th All-Union Conference on Probability Theory and Mathematical Statistics, Vilnius, 1962, pp. 309-323.

- 125.* Belyaev, Yu. K., Gnedenko, B. V., and Kovalenko, I. N., "Basic Trends in Queueing Theory", Transactions of the 6th All-Union Conference on Probability Theory and Mathematical Statistics, Vilnius, 1962, pp. 341-355.
- 126.* Gnedenko, B. V., "On Non-loaded Duplication", *Technical Cybernetics*, No. 4, pp. 3-12 (1964).
- 127.* Gnedenko, B. V., "On Duplication with Renewal", *Technical Cybernetics*, No. 5, pp. 111-118 (1964).
- 128.* Grigelionis, B., "On the Convergence of Sums of Step Stochastic Processes to a Poisson Process", *Probability Theory and Its Applications*, Vol. 8, Issue 2, pp. 189-194 (1963).
129. Kendall, D., "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Imbedded Markov Chains", A Collection of Translations "Matematika", Vol. 3, No. 6, pp. 97-111, 1959.
130. Kendall, D. G., "Some Recent Works and Further Problems in the Theory of Queues", *Probability Theory and Its Applications*, Vol. 9, Issue 1, pp. 3-15, 1964.
- 131.* Khinchin, A. Ya., "The Mathematical Theory of a Stationary Queue", *Mat. Sbornik*, Vol. 39, No. 4, 73-84 (1932); see also Khinchin, A. Ya., *Studies in Queueing Theory*, Fizmatgiz, 1963.
132. Kiefer and Wolfowitz, "On the Theory of Queues with Many Servers", *Trans. Am. Math. Soc.*, 78, 1-18 (1955).
- 133.* Klimov, G. P., "Extremal Problems in the Theory of Queues", *Sbornik Cybernetics at the Service of Communism*, Vol. 2, Izd. "Energiya", pp. 310-325, 1964.
- 134.* Kovalenko, I. N., "Some Problems of Queueing Theory with Restrictions", *Probability Theory and Its Applications*, Vol. 6, No. 1, pp. 222-228, 1961.
- 135.* Kovalenko, I. N., "Certain Analytical Methods in Queueing Theory", *Sbornik Cybernetics at the Service of Communism*, Vol. 2, Izd. "Energiya", pp. 325-338, 1964.
136. Lindley, "The Theory of Queues with a Single Server", *Proc. Cambridge Phil. Soc.*, Vol. 48, 277-289 (1952).
137. Maryanovich, T. P., "Reliability of Systems with Loaded Reserve", *Doklady Akad. Nauk Ukr. S.S.R.*, No. 8, 964-967 (1961) (In Ukrainian).
138. Miller, R. G., "Priority Queues", *Ann. Math. Stat.*, Vol. 31, No. 1, 86-106 (1960).
- 139.* Ososkov, G. A., "A Limit Theorem for Flows of Homogeneous Events", *Probability Theory and Its Applications*, Vol. 1, No. 2, pp. 274-282, 1956.
- 140.* Sevastyanov, B. A., "An Ergodic Theorem for Markov Processes and Its Application to Telephone Systems with Refusals", *Probability Theory and Its Applications*, Vol. 2, Issue 1, 1957.
- 141.* Shakhbazov, A. A. and Samandarov, E. G., "On Servicing a Non-ordinary Flow", *Sbornik Cybernetics at the Service of Communism*, Vol. 2, Izd. "Energiya", pp. 338-353, 1964.
142. Smith, W. L., "Renewal Theory and Its Ramifications", *J. Roy. Stat. Soc.*, Ser. B., Vol. 20, 1958.
- 143.* Solov'yev, A. D., "On Stand-by Systems Without Renewal", *Sbornik Cybernetics at the Service of Communism*, Vol. 2, 83-122, Izd. "Energiya", 1964.
144. Takács, L., "Some Probability Problems in Telephone Traffic", *Acta Math. Acad. Sci. Hung.*, Vol. 8 (1957) (In Hungarian).

SUBJECT INDEX

- Aftereffect, absence of 291, 294, 296, 297
- Algebra of events, σ -46
- Ars conjectandi* 206
- Averages, spatial 338
- Axiom of addition 47
 - extended 48, 49, 50, 129
- Axiom of continuity 49, 50
- Axioms, definition of 45
 - Kolmogorov's 48
- Axioms of probability 47-51
 - incompleteness of 48

- Banach's match box problem 105
- Bayes's theorem 57, 301
- Bernoulli's formula 288
- Bernstein's theorem 218
- Bertrand's paradox 33, 44
- Binomial distribution 72
- Birkhoff-Khinchin ergodic theorem 338
- Birth and death processes 346, 350
- Boltzmann statistics 28
- Borel field 46, 47
- Bose-Einstein statistics 28
- Brownian motion 101, 106, 288, 297, 318, 323
- Buffon's needle problem 35, 38

- Canonical representation (of normal law and Poisson law) 274
- Chain, Markov (*see* Markov chain) 205, 291, 302
- Coefficient, correlation 174, 175, 323
 - diffusion 289
- Collectives (von Mises) 43
- Condition, compatibility 324
 - Lindeberg's 253, 254, 255, 258, 259, 284
 - Lyapunov's 259, 267
 - Markov 205
 - self-compatibility 326
 - symmetry 324
- Constant, Euler's 350
- Convergence in measure 210
- Convergence in probability 210
- Convergence of a sequence of random variables 209-212
- Convergence to the normal and Poisson laws, conditions for 282
- Convolution 274
- Correlation theory 326
- Covariance 174
- Cumulants 191, 222
- Curve, Cantor 133

- Deciles 191
- Decomposition of stationary processes, spectral 331, 334
- Degree of certainty of observer 16, 18
- Density of probability distribution 132
- Deviation, standard 180
- Die 20
- Difference (of events) 19
- Diffusion of gases 319
- Dispersion 169
 - technical 300
- Distribution, Bernoulli 260
 - Boltzmann 28
 - chi-square (χ^2) 148, 266
 - function of probabilities 125
- Distribution, function of a random variable 125, 127, 129, 130
 - gamma 375
 - Laplace 192, 285, 286
 - lattice 260
 - Maxwell 192
 - n -dimensional normal 247
 - nondegenerate (proper) n -dimensional normal 247
 - Pascal 191

- Poisson 98, 260, 364, 365
- Polya 194
- Student's 152
- Weibull 375

- Ellipses of equal probabilities 139
- Encounter problem 46
- Equation, Fokker-Planck 290
 - generalized Markov 302, 303, 304, 307, 313
 - Kolmogorov's first 304
 - Kolmogorov's second 306
 - Markov 313
- Equations, Kolmogorov's 303-311
 - Kolmogorov-Feller 311, 312, 318
- Essai philosophique sur les probabilités* (Laplace) 39
- Event, certain 13, 20, 21, 47
 - decomposable into particular events 21
 - elementary 59
 - impossible 13, 20, 21, 47
 - random 12, 21, 45, 46
 - laws of 22
 - simple 61
 - sure 12, 21
- Events, collectively dependent 55
 - collectively independent 54
 - complementary 47
 - complete group of 21
 - complete group of pairwise mutually exclusive 21
 - contrary 20, 47
 - elementary 21, 45
 - equivalent 19
 - field of 21, 22, 46
 - field of, Borel 46
 - mutually exclusive 20, 47
 - simple 21, 22, 45
- Existence of limiting values (von Mises) 43
- Expectation, mathematical 164-169, 176, 190
 - theorems on 176-182
- Expectation, mathematical, defined in the axiomatics of Kolmogorov 182-185
- Expectation of a constant 176
- Expectation of a product 178
- Expectation of a sum 176
- Expectation, sign of 179

- Flow, elementary 341
- Formula, Bayes' 298, 301
 - inversion 224, 227, 229
- Kolmogorov's 285
- Stirling's 75
- Taylor's 305, 307
 - of total probability 55, 299
- Formulas, Erlang's 345, 352
- Formulas of Bayes 57
- Frequency and probability 38
- Function, Borel 128
 - conditional density 300
 - conditional distribution 134
 - correlation 326, 329
 - distribution 124, 125, 127, 129, 130, 189, 190
 - distribution (of a quotient) 150
 - distribution (of a random variable) 125
 - distribution (of a sum) 143
 - distribution, n -dimensional (of a random vector) 134
 - probability density 132, 138
- Functions, characteristic 219-250
 - characteristic (of multidimensional random variables) 243-248
 - conditional distribution 298
 - jump 322
 - multidimensional distribution 134-142
 - positive definite 239-242
 - of random variables 142-155

- Games of chance 7
- Geometry, non-Euclidean 8

- Implication 18
- Independence of events 53
- Inequality, Bunyakovsky-Cauchy 171
 - Cauchy-Bunyakovsky-Schwarz 327
- Inequality, Chebyshev's 198, 199, 203, 204, 205, 212, 213, 327
 - Kolmogorov's 213, 215
 - Schwarz 171
- Integral, Lebesgue 169, 182, 183
 - Poisson 132
 - s -special 335, 336
 - Stieltjes 143
 - defined 155-160, 169, 273, 299, 332
 - stochastic 331

- Law, associative 22
 - of binomial probability distribution 72
 - Cauchy's 166, 171
 - commutative 22

- of the excluded middle 17
 - of conservation of matter 13
 - distribution 130
 - of distribution, normal 167, 172
 - distributive 22
 - idempotency 22
 - of large numbers 92, 195-218
 - Chebyshev's form of 198-206
 - mass-scale phenomena and the 195, 197
 - a necessary and sufficient condition for 206
 - strong 209-218
 - logarithmic normal distribution 193
 - Maxwell 149
 - normal 268, 273, 282, 284, 320
 - normal distribution 139
 - Pascal's 198
 - Poisson 98, 126, 167, 178, 228, 268, 273, 274, 282, 284, 298, 316, 320
 - Simpson distribution 145
 - Student's 153
 - two-dimensional normal 175
- Laws, infinitely divisible distribution, canonical representation of 270
 - definition of 268
 - theory of 267
- Length of an interval 336
- Likelihood, equal 18
- Loss of a call 340

- Markov chain, definition of 107-123, 181
- Markov chain, simple 107
- Mass-scale operation 13
- Mass-scale phenomenon 10, 11
- Mathematical Methods of Statistics* (Cramér) 40
- Matrix, covariance 173
 - of transition probabilities 108
 - transition 108
- Mean 190
- Mechanics, statistical 324
- Median of a distribution 190
- Mode of a distribution 191
- Molecular speeds 319
- Moment about the origin, κ th 185
- Moment, absolute 186, 188, 190
 - central 185
 - of the κ th order 185
- Moments 185-191
 - mixed central (of the second order) 173
 - problems of 190
- Mortality tables 63
- Motion, Brownian 101, 103, 288, 297, 318, 323

- Ordinariness 292, 294, 297
- Outcome, possible 23

- Paradox, Bertrand's 33, 34, 36, 44
- Paradox of de Méré 67
- Period of a state 112
- Points, integral 89
- Probabilistic judgements 14
- Probabilistic regularities 13, 18
- Probability, axiomatic definition of 45
 - of causes, Bayes' rule for 57
 - classical definition of 16, 18, 22, 25
 - conception of (von Mises) 43
 - conditional 51, 52, 53, 56, 299
 - of congestion 342
 - definition of 47, 50
 - different approaches to definition of 15, 16
 - of an event 25
 - geometrical 33
 - of hypotheses, Bayes' formulas for 58
 - mathematical 14, 15, 16, 17, 18, 42
 - statistical 41, 42
 - statistical definition of 16, 32
- Probability, theory of 7, 11, 13, 14, 16, 17
 - theory of, axiomatic construction of 44-66
 - theory of, fundamental concept of (independence of events) 140
 - total, formula of 56
 - transition 108
 - unconditional 51
- Probability, Statistics and Truth* (von Mises) 44
- Problem, Banach's match box 105
 - Buffon's needle 35, 38
 - encounter 32, (in production) 33, 46
 - Erlang's 346
 - of limit theorems for sums, statement of 278
- Process, birth 346, 350
 - without aftereffect 291, 302, 312, 323, 324
 - continuous stationary stochastic 326

- continuous stochastic 303-311
- death 346
- Markov 290, 291, 343
- normal stochastic 328
- Poisson 291-298, 318, 341, 346
- Poisson (with leading function) 363
- probabilistic 287, 288
- purely discontinuous stochastic 311-318, (definition) 311
- random 287
- stationary 290, 291
- stationary stochastic 323
- step 362
- stochastic 287, 288, (definition) 291, 295, 302
- Processes, birth and death 346-355
 - with a discrete spectrum 330
 - with independent increments, homogeneous stochastic 318-323
 - distribution function of 320
 - Markovian 323
 - stationary 324
 - theory of 326
 - Wiener 318
- Product (of events) 19
- Quantile of order p , distribution 191
- Queueing system, single-server 355-361
- Queueing theory 9, 339-375
- Radioactive disintegration 319
- Random, at (explained) 35 and 37
- Random event 11, 12
- Random variable 8, 124, 125, 127
 - normally distributed 126
- Randomness (von Mises) 43
- Ranging fire, Bayes' formula in the theory of 58
- Rank of an interval 336
- Realization (of a stochastic process) 291
- Reliability theory 9
- Relief, loaded 368
 - nonloaded 368
 - partially loaded 367
 - stand-by (without repair) 349
- Scheme, Bernoulli 70, 71, 75, 91, 101, 126, 181, 216
 - Poisson 216
- Semi-invariants 191, 222
- Sequence of random variables, stationary 330
- Series, Maclaurin 238
- Service system with a waiting line (queue) 352
- Set of conditions 12
- Sets, Borel 49, 127, 131, 136
- Sets, Borel, fields of 49
- Space, n -dimensional Euclidean 135
 - probability 50
 - sample 18, 21, 69
 - of simple events 21
- Span, distribution 261
 - of a distribution 261
- Stability of frequencies 39
- Stand-by systems, theory of 367
- State, essential 111, 112
 - transient 111
 - unessential 111, 112
- States, communicating 111
- Stationary 291, 294, 337
- Statistics, Boltzmann 28
 - Bose-Einstein 28, 29
 - Fermi-Dirac 28, 30
 - of population 39
- Stochastic processes, theory of 287-338
- Stochastic regularities 13
- Sum (of events) 19
- System, single-server (one-channel) 355
- Theorem, Bayes' 57
 - of addition of probabilities 24
 - Bernoulli's 92, 164, 199, 206, 209
- Theorem, Bernstein's 218
- Bochner-Khinchin 240, 249, 328
- Borel's 209, 213, 216
- Chebyshev's 199, 200, 203, 205
- classical limit 251-266
- converse limit 236, 239
- Cramér's 148
- DeMoivre-Laplace 71, 116, 251, 278
 - of DeMoivre-Laplace integral 84, 90, 91
 - applications of 91, 120, 239
 - of DeMoivre-Laplace, local 75, 78, 83, 86, 96, 103, 104, 120
 - direct limit 235
 - ergodic 116
 - ergodic (Birkhoff-Khinchin) 334-338
 - Feller's 347
 - Grigelionis 364
 - Helly's first 231-232, 236, 272, 276, 277

- Helly's generalized second 234, 235
- Helly's second 232-234, 236, 272, 275, 276
- integral limit 84, 88
- Khinchin's 203, 218, 327, 334
- Khinchin's (on the correlation coefficient) 323
- Kolmogorov's 215
- Lagrange 301
- Laplace 88, 201
- Lebesgue 230
- limit (for flows) 361
- limit (for infinitely divisible laws) 275-278
- on limiting probabilities 113
- local 84
- local Laplace 75
- local limit 75 *et seq.*, 259-266
- Lyapunov's 254-259, 266, 278
- Markov's 205, 206, 218
- Theorem, multiplication 53, 54, 56, 296
 - Poisson's 96-101, 201, 285
 - Slutsky's 333
 - uniqueness 224, 227, 228, 238
- Theorems, Helly's 230-235
 - limit (for characteristic functions) 235-239
 - limit (for sums) 279-282
- Theory of errors 7
- Theory of Markov chains 107
- Theory of operators, spectral 334
- Theory of stochastic (probabilistic, random) processes 8
- Traffic, incoming 341
- Transformations, Fourier 219
- Transforms, Laplace 369, 370, 372
- Trial 23
 - definition of 70
- Trials, independent 69, (definition) 70
- Truncation, method of 203
- Value, principal (of a logarithm) 256
- Variable, lattice random 262
 - multidimensional random 134
 - n -dimensional random 134
 - one-dimensional random 136
 - random 125, 126, 127
- Variables, continuous 132
 - discrete 131
 - uncorrelated random 329
- Variance 169-175, (defined) 169, 173, 175
 - of a constant 179
 - of a sum 179, 180
 - theorems on 179, 180
- Vector, random 134
 - uniformly distributed random 136
- Venn diagram 19, 21